

AFIT/GOR/ENG/97M-04

Decision Boundary Analysis Feature Selection  
for Breast Cancer Diagnosis

THESIS  
Daniel West Gregg  
Captain, USAF

AFIT/GOR/ENG/97M-04

19970402 075

Approved for public release; distribution unlimited

**DTIC QUALITY INSPECTED 1**

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U. S. Government.

AFIT/GOR/ENG/97M-04

Decision Boundary Analysis Feature Selection  
for Breast Cancer Diagnosis

THESIS

Presented to the Faculty of the School of Engineering  
of the Air Force Institute of Technology

Air University

In Partial Fulfillment of the  
Requirements for the Degree of  
Master of Science, Operations Research

Daniel West Gregg, B.S., Physics  
Captain, USAF

March, 1997

Approved for public release; distribution unlimited

## THESIS APPROVAL

Student: Daniel W. Gregg, Capt, USAF      Class: GOR-97M

Title: Decision Boundary Analysis Feature Selection for Breast Cancer Diagnosis

Defense Date: 20 February 1997

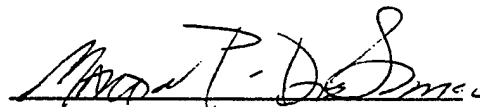
Committee:    Name/Title/Department

Signature


Advisor      Steven K. Rogers  
Professor  
Department of Electrical Engineering



Reader      Martin P. DeSimio  
Assistant Professor  
Department of Electrical Engineering



Reader      Kenneth W. Bauer  
Professor  
Department of Operational Sciences



## *Acknowledgements*

In every realm of the human experience, there is a beginning and there is an end. The end is with us and I would like to give thanks to the many people who have made this endeavor bearable. To my "family", Neal Bruegger, José Belano, Christine Davis, Angie Giddings, Sonia Leach, Wen-Chieh Liu, David Lyle and Heath Rushing, I cannot say enough. Your love and kindness have been immeasurable. Your impact eternal. I thank you all for being you and for loving me.

I would also like to thank Dr. Steven Rogers for creating a wonderful learning environment. Studying under Dr. Rogers has truly been a pleasure. In addition, I would like to thank Dr. Desimio and Dr. Bauer for their guidance and suggestions. I would like to specifically acknowledge Dr. Bauer's input concerning this thesis. Your remarks and suggestions are much appreciated and heeded. With great respect, I thank you all. Finally, I would like to acknowledge the beginning and the end, apart from which we can do nothing.

Daniel West Gregg

## *Table of Contents*

	Page
Acknowledgements . . . . .	iii
List of Figures . . . . .	vii
List of Tables . . . . .	viii
Abstract . . . . .	ix
 I. Introduction . . . . .	 1-1
1.1 Background . . . . .	1-1
1.2 Problem Statement . . . . .	1-3
1.3 Scope . . . . .	1-3
1.4 Overview . . . . .	1-3
 II. Theory . . . . .	 2-1
2.1 Pattern Recognition . . . . .	2-1
2.2 Computational Expense . . . . .	2-3
2.3 Theoretical Limitations . . . . .	2-3
2.4 Classification Accuracy . . . . .	2-4
2.5 Feature Saliency Techniques . . . . .	2-4
2.5.1 Classical Techniques . . . . .	2-5
2.5.2 Modern Techniques . . . . .	2-13
2.6 Comments . . . . .	2-23
2.7 Conclusion . . . . .	2-24

	Page
III. Methodology . . . . .	3-1
3.1 Correlation Analysis Feature Screening . . . . .	3-2
3.2 Cascade-Correlation on the Input Features . . . . .	3-3
3.3 Classifier-Based Saliency . . . . .	3-4
3.3.1 Ruck Saliency at Data Points . . . . .	3-4
3.3.2 Ruck Saliency at the Decision Boundary . . . . .	3-5
3.4 Summary . . . . .	3-6
IV. Analysis and Results . . . . .	4-1
4.1 Database . . . . .	4-1
4.2 Initial Screening . . . . .	4-1
4.3 Initial Classification . . . . .	4-3
4.3.1 Top 20 Uncorrelated Features . . . . .	4-5
4.4 Classifier-Based Feature Saliency . . . . .	4-6
4.5 Comparative Analysis of Ruck Variants . . . . .	4-8
4.6 Summary . . . . .	4-9
V. Conclusions . . . . .	5-1
5.1 Violation of Foley's Rule . . . . .	5-1
5.2 Fisher's Discriminant Ratio . . . . .	5-1
5.3 Correlation Analysis . . . . .	5-2
5.4 Cascade Correlation Analysis . . . . .	5-2
5.5 Known Data vs. Decision Boundary . . . . .	5-3
Appendix A. Ruck at the Decision Boundary vs. EDBFA . . . . .	A-1
Appendix B. Spectral Decomposition . . . . .	B-1
Appendix C. Ranking Equivalency . . . . .	C-1

	Page
Appendix D.      Ruck at the Decision Boundary vs. Ruck at Known Data . . .	D-1
Appendix E.      Should Cascade Correlation Work? . . . . .	E-1
Bibliography . . . . .	BIB-1
Vita . . . . .	VITA-1



## *List of Figures*

Figure	Page
2.1. The pattern recognition process . . . . .	2-1
2.2. Sample distribution of feature 1 for both classes . . . . .	2-5
2.3. Sample distribution of feature 2 for both classes . . . . .	2-6
2.4. Data cluster . . . . .	2-9
2.5. MLP architecture . . . . .	2-14
2.6. Decision boundary with unit normals . . . . .	2-19
2.7. Unit normals relative to origin . . . . .	2-20
4.1. Correlations of features with class . . . . .	4-2
4.2. Correlation matrix . . . . .	4-5
4.3. Saliency flow chart . . . . .	4-7
A.1. Decision boundary for two dimensional two class problem . . . . .	A-1
D.1. Two dimensional two-class problem . . . . .	D-1
D.2. Neural network response surface . . . . .	D-2
D.3. Two class pattern distributions . . . . .	D-3
D.4. Poorly trained network . . . . .	D-4
D.5. Properly trained network . . . . .	D-4
E.1. Highly correlated features . . . . .	E-1

## *List of Tables*

Table	Page
2.1. Exemplars . . . . .	2-8
4.1. Features by type . . . . .	4-1
4.2. Top 20 features from initial screening . . . . .	4-2
4.3. Classification results using top 10 correlation features . . . . .	4-3
4.4. Classification results using top 10 Fisher features . . . . .	4-4
4.5. Classification results using ten random features . . . . .	4-4
4.6. Pool of 20 features using cascade correlation . . . . .	4-6
4.7. Top 10 features by frequency of occurrence . . . . .	4-7
4.8. Classification results using top 10 features from combined techniques . . .	4-7
4.9. Top 10 features using two variants of Ruck's saliency . . . . .	4-9
4.10. Classification results using top 10 features from Ruck's at data points . .	4-9
4.11. Classification results using top 10 features from Ruck's at decision boundary	4-9
A.1. Normal vectors . . . . .	A-2
A.2. Ruck at the decision boundary . . . . .	A-3
D.1. Ruck saliency values . . . . .	D-2
D.2. Ruck saliency with poorly trained network . . . . .	D-4
D.3. Ruck saliency with properly trained network . . . . .	D-4

*Abstract*

The general pattern recognition problem always involves the extraction of features to be used in pattern classification. There are no theoretical limitations to the number of features which can be obtained for a given pattern recognition problem. There are however, many practical concerns which compel the researcher to reduce the feature space dimensionality to a set of most salient features. This process, called dimensionality reduction, has been a thoroughly researched area of pattern recognition. This thesis has a three-fold contribution. First, a comparison will be made between Ruck's saliency as proposed in previous works and a new variant of Ruck's saliency to be introduced. The results of the new method will prove to be superior. Classification accuracy is improved by over 7 percentage points. Secondly, a proposal will be presented which establishes how one may use the eigenvalue/eigenvector pairs from DBFA [22] for feature saliency. This proposal will also provide proof of the equivalence of DBFA and the Ruck variant proposed in this thesis. Because the Ruck variant is easier to calculate, it is suggested that DBFA is unnecessary. Finally, this thesis will investigate the application of classifier-free feature screening of a large feature space. A correlation-based procedure will be developed which has proven to outperform other saliency metrics such as the Fisher ratio and derivative-based techniques such as Ruck's saliency. This procedure has produced classification accuracies 10 percentage points higher than that of Fisher saliency while achieving a slightly better (2 percentage points) classification accuracy than the best derivative-based results. In addition, a combined process will be implemented which is superior to any stand alone technique. Classification results from the combined technique are 10 percentage points higher than the best results from any of the other methods. The applicability of the proposed technique is limited in this research to two-class pattern recognition problems, but may be extended to multi-class problems.

# Decision Boundary Analysis Feature Selection for Breast Cancer Diagnosis

## *I. Introduction*

The National Cancer Institute (NCI) estimated that in 1994 182,000 women were newly diagnosed with breast cancer with over 46,000 deaths per year [25]. Assuming a longevity of 79 years, the lifetime risk that a woman will develop breast cancer is 1 in 8. Breast cancer is the second leading cause of cancer death among women and accounts for 4 percent of all deaths of women in North America each year.

Early breast cancer diagnosis and detection continues to be a focus of researchers around the world. Early detection is vital for patient survivability. Although screen/film mammography has been used for many years and is currently the best method for breast cancer screening, 10 percent of breast cancers do not show up on these X-rays [25]. The most sobering statistic concerning breast cancer diagnosis is that 10 to 30 percent of negative diagnoses are later determined to be cancerous [25]. In two thirds of these false negatives, the malignant region was evident upon reexamination of the mammogram [16].

### *1.1 Background*

There are three major areas in which researchers have focused in order to improve the breast cancer diagnosis process. One research focus is the area of image acquisition. Ultrasound and Infra-red techniques exist but X-ray mammography is currently the most common means of imaging for breast cancer screening. The limitations of this are numerous. As stated before, 10 percent of all breast cancers will not be evident in an X-ray. The reason is that there are very subtle differences between normal and cancerous breast tissue. X-ray mammography may not be able to sufficiently contrast these regions. In particular,

denser breast tissue in younger patients obscures abnormal tissue. These image acquisition limitations are being addressed by such groups as MedDetect, Fischer Imaging, Nova R&D and ThermoTrex Corp. [30]. Most of this activity utilizes imaging technology reaped from military research in the area of target recognition.

The second major focus deals not with imaging but rather image processing [9, 18, 4, 10]. Historically there has been little image processing in the field. Typically, radiologists simply view the film mammograms as given. More recently, the plethora of military image processing techniques have been applied to breast cancer screening [26, 6]. It is now commonplace for clinicians to digitize X-ray films. This usually provides some contrast enhancement but more importantly it allows the researcher to extract information from the image using a computer. This information may not be readily apparent to the unaided eye of the radiologist. Extracting information from an image is commonly referred to as feature extraction. There can be hundreds of features in an image, not necessarily visible, which may indicate abnormality of tissue. Although this is an enormous area of research, there is currently no implementation in the field of computer aided diagnosis.

The final area of intense research is in actual diagnosis [5, 28, 12]. Researchers feel strongly that given relevant features from image processing techniques, they can correctly classify regions of interest in a breast image as malignant or benign. The key is to find relevant features. As was stated previously, in any pattern recognition problem it is always possible to extract nearly countless features from the pattern image. Examples could be Fourier coefficients, biorthogonal wavelet coefficients, contrast measures, entropy, angular second moment and eigenmass [5, 10]. Many of these features are *ad hoc* metrics that we somehow feel may be important in determining if a given region is malignant or benign. Most researchers would agree that if cancer is in the image, we should be able to find it. The problem is simply that we don't know what indicates cancerous tissue. Typically then, many, sometimes hundreds of features are extracted in hopes of finding the salient or most important ones for classification. Unfortunately, when classifying data, the computational costs are huge if the number of features is large [31]. In addition, there is a theoretical

limit in the number of features a classifier can use based on the number of training samples available [14]. Due to the theoretical and computational limitations [1], it is almost always necessary to reduce the beginning feature set to some subset of the most salient features.

Current research at the Air Force Institute of Technology involves all three of the aforementioned interest areas. Many researchers over the years have contributed to the pool of features which can be extracted from a digitized mammogram. The current feature count is 170. In previous research thrusts, feature saliency techniques have been applied to smaller sets of features [5]. In each of these contributions, the most salient features from each of these subsets were identified. To date, no one has identified the most salient features from the entire set. The reason is that most saliency metrics require a non-parametric classifier such as an MLP. To train an MLP using 170 features, would require a huge database of training exemplars to avoid violation of Foley's rule [14]. The computations would also be very time consuming.

### *1.2 Problem Statement*

This research will develop a correlation procedure for screening a large feature set without the use of a trained classifier. The results will be compared to established saliency metrics such as the Fisher ratio and derivative-based techniques such as Ruck's saliency.

### *1.3 Scope*

This research will develop a procedure for screening a large feature set without the use of a trained classifier. In addition, a variant of the Ruck saliency metric will be introduced. Theoretical comparisons will be made between the results of this research and those of classifier-based feature saliency metrics such as Ruck's saliency at the data points [34].

### *1.4 Overview*

With the motivation and the problem having been stated here in Chapter I, Chapter II will present related research and background information. Chapter III will discuss the specific

methodology used in development of a classifier free screening technique. Data description and analysis will be discussed in Chapter IV. Conclusions and research suggestions will be presented in Chapter V.

## II. Theory

### 2.1 Pattern Recognition

Pattern Recognition can be mathematically defined as a mapping from  $R^n \rightarrow R^m$ ,  $m \leq n$ . To be more specific, in pattern recognition we are typically concerned with assigning a vector in  $R^n$  to one of  $m$  classes. As an example, consider a vector of human attributes such as age, body fat, blood pressure, shoe size, hair length and skin color. This produces a 6-dimensional vector for each human. In pattern recognition, these attributes are called features and the vector is called the feature vector. Now, given a feature vector for any randomly chosen individual, we wish to classify that person as male or female based solely on the feature vector. This is the essence of pattern recognition. The complete pattern recognition problem involves three steps and is illustrated in Figure 2.1. The first step involves

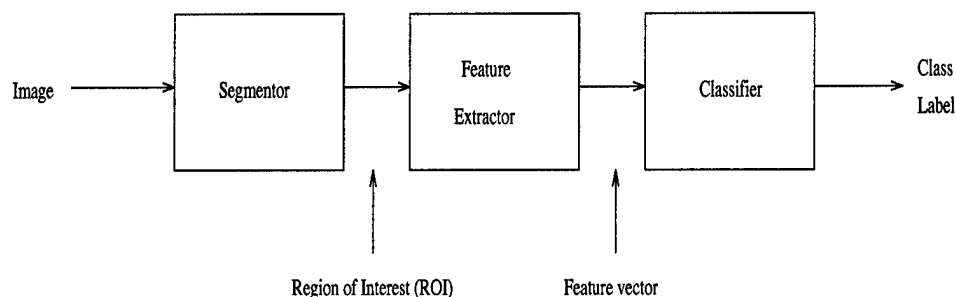


Figure 2.1 The pattern recognition process

the segmentation of the measurement space. For example, given a satellite photo covering an area of enemy occupation, the segmentation procedure will determine which regions of interest (ROI's) of the photo may contain possible targets. This is accomplished using some computer aided filtering scheme. Given these ROI's, the next step extracts certain attributes or features from the region. Here, the features could be coefficients from a Fourier transform or any of countless other features. The features are assumed to be useful in determining the target characteristics. In the final step, a vector of extracted features for a given ROI is used to classify the suspect region into one of a number of target classes. The classification



success is highly dependent on the saliency of the chosen features.

Returning to the gender example, it is instructive to note the correlation between age and shoe size. Certainly, age or shoe size alone have no discriminatory power with regard to gender, but together they provide useful information. Also notice that skin color probably has no bearing on gender status. This is a simple instructive example which was intuitively easy to analyze. Unfortunately, in real world applications, the features may have extensive multi-collinearities and many may be useless in the discrimination task. In this case, they are never as easy to identify as in this example. The identification of these correlated features and useless features is one of the most important problems in pattern recognition and is the focus of this entire thesis.

The Air Force Institute of Technology has studied many pattern recognition problems [39, 5]. Traditionally, AFIT in conjunction with Wright Laboratories, has been concerned with the target recognition problem. Much work has been done at AFIT over the last 30 years regarding this particular pattern recognition problem. More recently, researchers at AFIT have formed a breast cancer research group with the goal of applying decades of target recognition technology to the related problem of breast cancer detection and diagnosis.

As previously discussed, one of the most important problems in pattern recognition is feature selection. It has been said many times before that good features make good classifiers. In any classification problem, it is almost certain that some features are more useful in discriminating between classes than others. These features which hold more information relative to the discrimination task are often termed salient features. As such, in any classification problem one seeks to find the most salient features from a given feature set.

Pattern recognition problems begin with the extraction of a large number of features from the measurement space. For example, the breast cancer research group at the Air Force Institute of Technology has developed 170 features which are routinely extracted from newly digitized mammograms. Although this is a large number, there is no reason researchers couldn't increase this number indefinitely as long as they felt the addition of a new feature may prove relevant to the classification problem. Unfortunately, there is no way to know *a*

*priori* whether a given feature is useful in the discrimination task. The high dimensionality of the feature space has been a fundamental problem in pattern recognition since its inception.

There are several theoretical and computational difficulties that arise from high dimensional feature spaces. The following three sections briefly discuss these difficulties.

## 2.2 *Computational Expense*

In the pattern recognition problem, the second phase involves the extraction of features from a given ROI. Each of these features is typically some numerical quantity or metric. Some of these could be the coefficients of the Fourier transform or possibly wavelet coefficients. In any case, the determination of a feature value involves numerical calculations, some of which may be quite complex. If the dimension of the feature space is high, the feature extraction process could be very time consuming.

When a non-parametric classifier such as a neural network is used in the classification phase, the training time is significantly affected by the dimensionality of the feature space. Typically, this is a one time initial cost and would not be of great concern. However, there are pattern recognition problems which require “on-the-fly” retraining [39]. In these instances, training times become very important and the dimensionality of the feature space is a significant factor.

## 2.3 *Theoretical Limitations*

In addition to the computational considerations, there are theoretical limitations surrounding the feature space dimensionality. The so called *curse of dimensionality* [1] states that as the number of input features increases, the number of feature vectors must increase exponentially for accurate classification. Other works such as Foley [14] give a more practical rule-of-thumb. By the Foley criteria,  $3n$  feature vectors should be used for each class. In the breast cancer diagnosis problem, we have a two class problem (*malignant*, *benign*)

with a 170 dimensional feature space ( $n = 170$ ). As such, we require 1020 feature vectors to meet the Foley criterion. It should be noted that the Foley criterion is actually a lower bound [31]. In many pattern recognition problems, acquisition of feature vectors is time consuming, costly or simply impractical. In these cases, the number of feature vectors is limited and the *curse of dimensionality* dictates that we must reduce the dimensionality of the feature space. In the breast cancer problem, we currently have approximately 70 ROI's for which we have truth data (*i.e. regions of interest which have been biopsied*). Although this is a fairly small number, it took several years to build this database.

#### 2.4 Classification Accuracy

Finally, there is one additional consideration. It has been shown that the use of insignificant features as input into a neural network may reduce classification accuracy [31]. In Rogers' paper an example was given using the breast cancer database and 21 orthogonal wavelet features. Initially, all 21 features were used to train the network. After employing feature saliency techniques, the feature space was reduced to just 7 features. With the feature space reduced by two thirds, the network proved to have a higher classification accuracy. This result has been seen many times in the literature [29, 3].

#### 2.5 Feature Saliency Techniques

The thrust of this thesis is on the development and comparison of feature saliency methods. This chapter presents much of the theory surrounding the most common techniques used in practice. Not all of the following methods will be analyzed for comparisons but they will be explained here because it is important to understand what assumptions surround the different techniques. The remainder of this chapter is divided into two sections. Section 2.5.1 provides the background theory for classical feature selection techniques. Section 2.5.2 presents more recent developments in feature saliency. Each of these sections is further divided into two subsections, classifier-free and classifier-based techniques.

*2.5.1 Classical Techniques.* The classical techniques are so named due to their longstanding prominence in the field of pattern recognition. The oldest of these dates back to 1936 [13]. Most of these require no classifier to perform the technique.

*2.5.1.1 Classifier-Free Feature Saliency.* One of the oldest and most fundamental measures used in feature selection is the Fisher Discriminant [13]. For the two class discrimination problem, the mathematical relation defining Fisher's Discriminant is

$$f = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \quad (2.1)$$

Here,  $\mu_1$  and  $\mu_2$  represent the mean of a given feature for class 1 and class 2 respectively and  $\sigma_1$  and  $\sigma_2$  are the standard deviations of the respective classes in the feature dimension. To illustrate, suppose we have a two class problem and we wish to evaluate the discriminatory effectiveness of two different features. Figure 2.2 shows the sample distributions of feature 1 for class 1 and class 2. The mean for each class is marked with vertical lines. Notice the

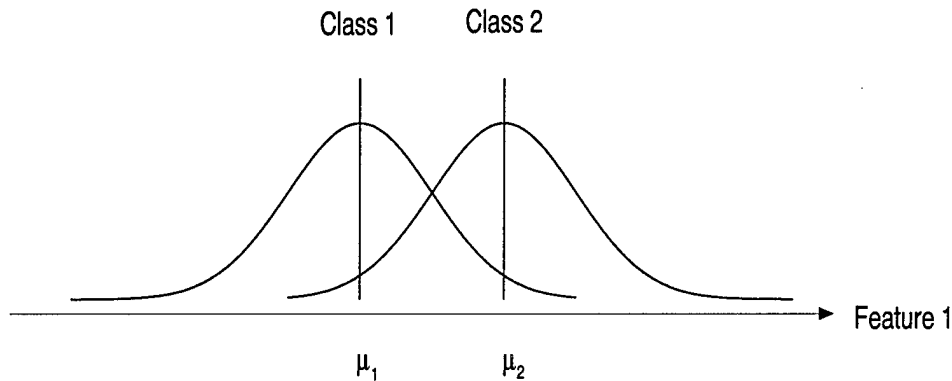


Figure 2.2 Sample distribution of feature 1 for both classes

means are well separated but there is significant overlap due to the large variances of the distributions. Now consider the sample distributions of feature 2 for each class. These are shown in Figure 2.3. Notice here, the means are no longer as far apart but the variances are significantly reduced. There is little overlap. It should be intuitively clear that feature 2 would be better for discriminating between class 1 and class 2. The Fisher Discriminant

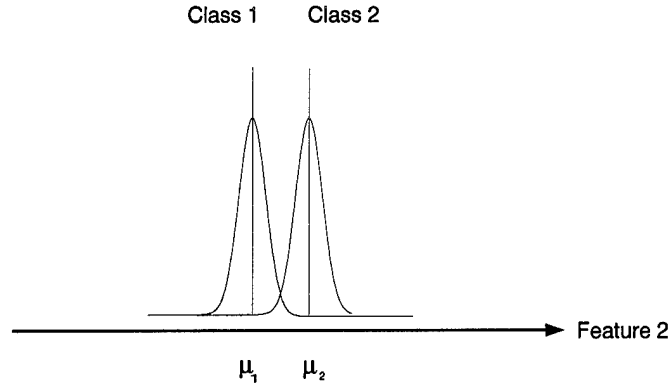


Figure 2.3 Sample distribution of feature 2 for both classes

of equation 2.1 quantifies this intuition. For feature 1,  $\mu_1 = 1$ ,  $\mu_2 = 2$  and  $\sigma_1 = \sigma_2 = .5$ . For feature 2,  $\mu_1 = 1$ ,  $\mu_2 = 1.5$  and  $\sigma_1 = \sigma_2 = \frac{1}{8}$ . The calculated  $f$  values are  $f = 2$  and  $f = 8$  for feature 1 and feature 2 respectively. It is important to realize that the Fisher Discriminant assumes these marginal distributions are Gaussian. The applicability of this measure is questionable for asymmetrical or multi-modal distributions.

Pattern recognition often involves more than two classes and we resort to a generalized Fisher Discriminant often called the F-ratio. Notice in equation 2.1, the expression is roughly the variance of the means over the mean of the variances. The generalized Fisher Discriminant or the F-ratio is given by [27]

$$F = \frac{\text{variance of the means(over all classes)}}{\text{mean of the variances(within-classes)}}$$

Given  $N$  samples for each of  $m$  classes, the ratio becomes

$$F = \frac{\frac{1}{(m-1)} \sum_{j=1}^m (\mu_j - \bar{\mu})^2}{\frac{1}{m(N-1)} \sum_{j=1}^m \sum_{i=1}^N (x_{ij} - \mu_j)^2} \quad (2.2)$$

where  $x_{ij}$  is the  $i$ th measurement from class  $j$ ,

$\mu_j$  is the mean of all measurements for class  $j$ ,

and  $\bar{\mu}$  is the grand mean.

It should be emphasized that the Fisher Discriminant or the generalized form evaluate a single feature. One possible approach may be to calculate the F-ratio for each feature, then select an appropriate number of features to retain from the ranked set. The danger in this approach is that the features have been considered in isolation. Interactions have not been taken into account. It is highly likely, as in the gender classification example, that some features may be correlated.

Fukunaga develops four additional measures which are further generalizations of the F-ratio concept [15].

1.  $J_1 = \text{tr}(S_w^{-1}S_b)$
2.  $J_2 = \ln|S_w^{-1}S_b|$
3.  $J_3 = \text{tr}S_b - \mu(\text{tr}S_w - c)$
4.  $J_4 = \frac{\text{tr}S_b}{\text{tr}S_w}$

where  $\text{tr}$  is the trace;  $S_w$  and  $S_b$  are respectively, the *within-class* and *between-class* scatter matrices defined and fully explained later in this section. They have the advantage of simultaneously considering a set of features rather than individually. The development of another feature saliency technique, the Karhunen-Loève transformation is a necessary prelude to the complete understanding of the generalized F-ratio and will be developed next.

*Discrete Karhunen-Loève Transformation.*

Suppose you are given

a distribution of exemplars in  $\mathbf{R}^n$ . Finding the covariance matrix of this distribution will most likely reveal off-diagonal elements which are non-zero. As such, these features are correlated. Techniques employed to remove this correlation are referred to as diagonalization or canonical analysis. The concept of these transformations involves rotation of the feature space such that the new feature axes align with the directions of maximum variance. This rotation can always be found. An example of this transformation is given next. Consider the data in Table 2.1. The data are plotted in Figure 2.4. One can immediately see the

feature 1	feature 2
3.8	4.6
3.5	4.3
3.4	3.7
3.4	3.5
3.8	3.3
4.5	3.6
4.9	4.2
5.2	4.1
5.9	4.9
6.2	5.4
6.3	5.9
5.7	6.2
4.9	6.0
4.3	5.7
3.6	5.0
4.4	4.1
5.3	4.8
6.0	5.7
5.9	6.2
4.6	5.5
4.5	4.9
5.1	5.5

Table 2.1 Exemplars

directions of maximum variance. It is also apparent that a rotation of approximately  $45^\circ$  will align the feature axes appropriately. The diagonalization procedure for determining the precise rotation is as follows. First, determine the covariance matrix of the exemplars from equation 2.3.

$$\Omega = \frac{1}{(N-1)} \sum_i^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \quad (2.3)$$

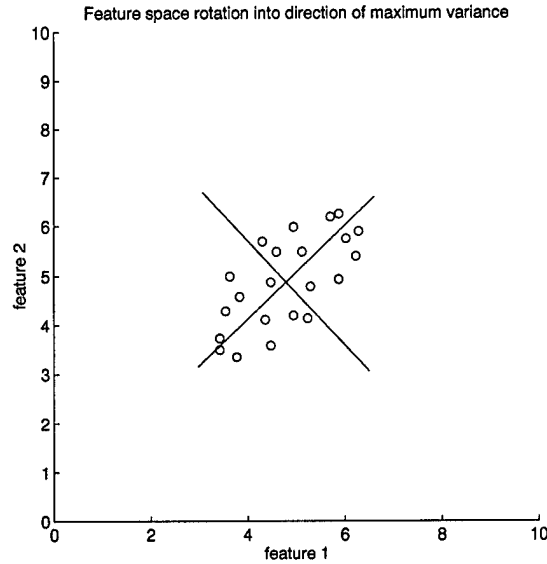


Figure 2.4 Data cluster

where  $\mathbf{x}_i$  is a given feature vector and  $\bar{\mathbf{x}}$  is the mean feature vector.  $i$  is the index over the entire set. Given the data in Table 2.1, the covariance matrix is shown below.

$$\Omega = \begin{bmatrix} .9011 & .5781 \\ .5781 & .8404 \end{bmatrix} \quad (2.4)$$

Notice  $\Omega$  is not diagonal, which is most common. In other words, the feature set is highly correlated. We wish to find a transformation  $\mathbf{A}$  such that  $\Omega$  is diagonalized. In short, we seek the matrix  $\mathbf{A}$  such that

$$\mathbf{A}^T \Omega \mathbf{A} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n) = \Lambda$$



It is well known that the transformation  $\mathbf{A}$  is formed by using the normalized eigenvectors of  $\Omega$  as the columns of  $\mathbf{A}$  [7]. The eigenvectors are computed and the matrix  $\mathbf{A}$  is

$$\mathbf{A} = \begin{bmatrix} -.7254 & .6883 \\ -.6883 & -.7254 \end{bmatrix}$$

The corresponding eigenvalues are found from the transformation

$$\Lambda = \mathbf{A}^T \Omega \mathbf{A} = \begin{bmatrix} 1.4497 & 0 \\ 0 & .2919 \end{bmatrix}$$

Notice  $\Lambda$  is the transformed covariance matrix  $\Omega$  and is now diagonalized. More importantly, the variances in the new transformed space correspond to the eigenvalues of the covariance matrix from the original space. Notice the relative magnitudes of the eigenvalues. In this particular example, the smallest eigenvalue still accounts for roughly 17 percent of the total variance. This indicates that both directions in the rotated space have significant variance. Let's assume for a moment that one eigenvalue was significantly larger than the other. What does this mean? It tells us only that the direction in the *transformed space* corresponding to the larger eigenvalue contains most of the variance. This does not mean that this direction is the most discriminantly informative. In fact, this transformation has nothing to do with separability. It may however, allow us to represent the data with fewer features while minimizing the representation error. In this sense, we may be able to identify and eliminate certain features which have little impact on the representation of the data. Although the

KL transformation yields no information about feature saliency, its simplicity makes it a popular method for dimensionality reduction.

*Maximizing Separability.* The method of the orthogonal transformation as introduced above can be extended to provide information concerning the directions of maximum class separability [15]. In the extension, linear transformations, not necessarily orthogonal, are applied to combinations of between-class and within-class covariance matrices. The results are the four criteria found in Fukunaga [15]. Useful explanations of these criteria are also found in Parsons [27]. These criteria quantify the saliency of sets of features with respect to separability. The application of these metrics and others is generally called discriminant analysis. A qualitative explanation follows. Details can be found in Parsons [27].

In order to proceed with the following discussion, it is necessary to define the between-class covariance matrix  $S_b$ , and the within-class covariance matrix  $S_w$ . As suggested, the between-class covariance matrix is the covariance matrix formed from the centroids of the individual classes about their grand mean. The within-class covariance matrix is the average of the covariance matrices of all classes. Forgetting the definitions for a moment and considering what they intuitively represent, we see that the within-class covariance matrix tells us how *wide* the individual classes are on average whereas the between-class matrix gives us some indication as to the distance between the classes. Consider the separability measure  $S_w^{-1}S_b$ . If in a given direction, the ratio is large, we would expect good separability in that

direction. The goal is to find the directions of the measure  $\mathbf{S}_w^{-1}\mathbf{S}_b$  which give the largest ratios. To that end, we simply find the eigenvectors and eigenvalues of  $\mathbf{S}_w^{-1}\mathbf{S}_b$ . The eigenvector corresponding to the largest eigenvalue will be the direction of maximum separability in the *transformed* space. Once again, it should be noted that we've only determined an optimum direction in the *transformed* space which is still a linear combination of the original features. Given these eigenvectors, the selection process would involve choosing the features which contribute the most to the dominant eigenvectors. This allows us to prune the features and is sometimes known as dimensionality reduction.

*2.5.1.2 Classifier-Based Feature Saliency.* The previous techniques measured the saliency of a feature or set of features without actually using the features in classification. It was pointed out that under certain conditions there are severe limitations to these techniques. One possible approach that could be employed to eliminate these limitations would be to consider all possible subsets of features. Measuring the classification performance of each subset would then allow one to rank order the best performing subsets of features. Obviously this would account for all interactions between features and would produce the optimal feature space for the given training set. However, one can see that this exhaustive enumeration must consider a prohibitively large number of subsets. For example, suppose we have 170 features and we wish to reduce this number to the 10 most salient. This produces  $\binom{170}{10} \approx 10^{15}$  possible subsets. For each of these subsets, classification of the data set must be accomplished. Obviously, this is not an option unless the original feature space is

small and we only wish to eliminate a few features. Sambur [33] and Goldstein [17] proposed algorithms which fall into the subset selection category. These algorithms, called the Knock-out and Add-on algorithms require a much smaller subset of features to be compared. These produce subsets in which all correlations have been taken into account. The algorithms are, however, suboptimal since they do not consider all possible subsets.

*2.5.2 Modern Techniques.* In recent years, many new methods of analyzing feature saliency have been developed. Almost all of these are classifier-based. At least one approach can be applied without the use of a classifier [21].

*2.5.2.1 Classifier-Based Feature Saliency.* All but one of the following techniques specifically require the use of a non-parametric classifier such as a neural network. Before proceeding with the discussion of these techniques, it will be beneficial to provide the fundamental background information concerning the Multilayer Perceptron (MLP).

The Multilayer Perceptron consists of interconnected processing units as depicted in Figure 2.5. The MLP performs a mapping from  $R^i \rightarrow R^k$ . In this example, the MLP accepts an input vector of dimension  $i$  and produces an output vector of dimension  $k$ . Each interconnection has associated with it a weight. The weights between the input layer and the hidden layer are called the input layer weights. Those between the hidden layer and the output layer are called the hidden layer weights. Notice that the input layer weights are denoted by  $w_{ij}^1$ , where  $j$  is the index to the hidden layer nodes and  $i$  is the index to the input layer nodes. Similarly, the weights in the hidden layer are denoted by  $w_{jk}^2$ , where  $k$  is the

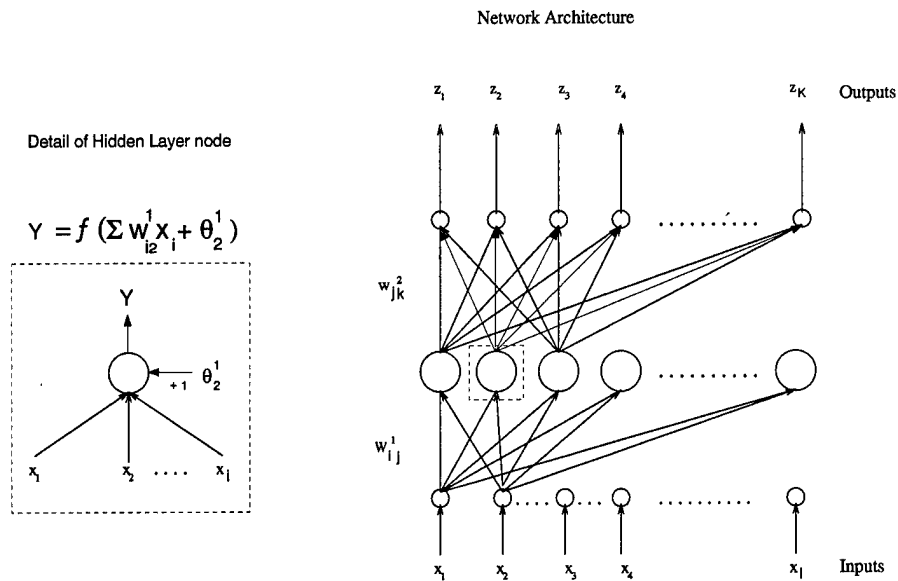


Figure 2.5 MLP architecture

index to the output layer nodes and  $j$  is the index to the hidden layer nodes. The output from any input layer node is simply the input. The output for any node in the hidden layer is denoted  $x_j^1$ . This output is computed from the weighted sum of the inputs to that node as shown in the detail of Figure 2.5. The weighted sum of the inputs is termed the activation. The function  $f$  is called the activation function and is typically one of the following.

$f(a)=a$  simple linear function

$f(a)=\tanh(a)$  hyperbolic tangent

$f(a)=\frac{1}{1+e^{-a}}$  sigmoid

The weights in an MLP are commonly determined through a training process called Backpropagation [32, 38]. In the Backpropagation algorithm (see [23] for details), an input is applied to the network and the associated output computed. The actual output is then compared with the desired output and the weights adjusted in an effort to minimize the

error between the desired and actual output. This process, called training, continues until all feature vectors in the data set have been processed enough times to reduce the overall error below some acceptable criteria. At this point the weights are fixed and the network is ready to classify new feature vectors.

The next three feature saliency techniques stem from the notion that the weights in a trained MLP encode all of the important information necessary for accurate classification. It is supposed that the relative magnitudes of the weights represent the relative importance of the input features [37].

*Ruck's Saliency.* Ruck's Saliency metric is a derivative-based technique which measures the sensitivity of an MLP's output to its input [11]. The metric is calculated from the partial derivatives of the outputs with respect to the inputs. The following equation formalizes this notion.

$$\Lambda_i = \sum_p \sum_m \sum_r \sum_k \left| \frac{\partial z_k}{\partial x_i}(\mathbf{x}_{m(r)}^p) \right| \quad (2.5)$$

here,  $k$  is the index over all outputs,  $m$  is the number of dimensions in the input space,  $r$  is the number of uniformly spaced points covering the range of each input feature and  $p$  is the index over all feature vectors in the training set.  $\mathbf{x}_{m(r)}^p$  is the vector at which the partial derivative is evaluated. It is located  $r$  units out in the  $m$ th dimension from the  $p$ th data point. The absolute value of the partial derivative is used because we are only concerned with the magnitude of the change in outputs. The sum over all outputs is necessary to measure the

full sensitivity to a change in input. In the computation of the partial derivative  $\partial z_k / \partial x_i$ , it is clear that there is a dependency on the weights of the MLP. Since the weights established after training an MLP depend on their initial values (which are randomly selected) and on the order of presentation of the data set, these weights are random variables. For this reason, we typically calculate the saliency in equation 2.5 for a number of distinct MLP's trained over the same data set but with different initial weights and presentation orders. With  $\Lambda_i^n$  representing the saliency of feature  $i$  calculated from the  $n$ th MLP, the average over  $N$  MLP's is given by

$$\Lambda_i = \frac{1}{N} \sum_{n=1}^N \Lambda_i^n \quad (2.6)$$

The subset of salient features can now be obtained from the ranked set of  $\Lambda_i$ .

*Tarr's Saliency.* Tarr's saliency metric is a weight-based method which does not require the calculation of derivatives [37]. Consequently, Tarr's metric is very easy to compute as shown in equation 2.7.

$$\tau_i = \sum_{j=1}^J (w_{ij}^1)^2 \quad (2.7)$$

where  $\tau_i$  is the Tarr saliency metric for feature  $i$ ,  $J$  is the number of hidden nodes and  $w_{ij}^1$  is the input layer weight between input node  $i$  and hidden node  $j$ . There are three variants of the Tarr metric. Equation 2.7 is the two norm. Another variant involves the sum of the magnitudes of the weights while the third takes the largest weight in magnitude as the saliency metric.

*Signal-to-Noise Ratio.* The previous techniques have proven extremely useful for determining the saliency *ranking* of a set of features. Unfortunately, these metrics only provide information concerning the saliency of one feature relative to another. It is then a matter of subjective opinion to select how many features to retain. To circumvent this problem, recent developments in feature saliency screening have been proposed. Belue and Bauer [2] have proposed a method in which a “noise” feature is added to the feature vectors of the training set of an MLP. The MLP is trained 10 to 30 times [2] and the average saliency for each feature including the “noise” feature is calculated. The “noise” feature saliency is presumed to be normally distributed so that an upper  $(1 - \alpha)\%$  confidence interval can be constructed for the average saliency of noise. The other features can then be compared directly to the useless noise feature. Those features whose mean saliency lie within the confidence interval of the noise feature, are likewise considered noise and can be removed from the feature set. This is very advantageous since it provides relative saliency information as well as a threshold for determining which features should be removed. In an effort to eliminate the need for training an MLP 10 to 30 times, the SNR saliency metric was proposed [19, 36]. The ratio can be calculated using any saliency metric but Bauer and Sumrell propose using the Tarr metric as shown in equation 2.8.

$$SNR_i = 10 \log \left( \frac{\sum_{j=1}^J (w_{ij}^1)^2}{\sum_{j=1}^J (w_{Nj}^1)^2} \right) \quad (2.8)$$

where  $SNR_i$  is the value of the metric for feature  $i$ ,  $J$  is the number of hidden nodes,  $w_{ij}^1$  is the input layer weight from node  $i$  to node  $j$  and  $w_{Nj}^1$  is the input layer weight from the



noise input node  $N$  to the hidden layer node  $j$ . The log transformation of the ratio converts the saliency metric to a decibel scale. The SNR screening method is given below [19].

1. Introduce a noise feature  $X_n \sim U(0,1)$  to the original set of features.
2. Normalize all features.
3. Begin training the MLP.
4. After each epoch, compute the SNR saliency measure for each input feature.
5. Interrupt training when the SNR saliency measures for all input features have stabilized.
6. Compute the classification error.
7. Identify the feature with the lowest SNR saliency measure and remove it from the set.
8. Continue training.
9. Repeat steps 4-7 until all features (except noise) have been removed.
10. Plot classification error against features removed.
11. Retain the first feature whose removal caused a significant increase in error as well as all subsequent features.

This approach has proven very effective in real world problems [20].

*Decision Boundary Feature Analysis.* Decision Boundary Feature Analysis (DBFA) [22] is a technique which goes straight to the heart of the classification problem. It is a method which conjectures that all of the relevant discriminatory information can be found from the decision boundary. This seems intuitively reasonable. By looking at the decision boundary for a given classification problem, we can usually identify the important discriminant directions. In Figure 2.6 for instance, it should be quite obvious that feature 2 plays the dominant role in classifying this data. It is true that feature 1 is necessary to correctly classify all data. However, if one were to classify using only feature 2, the results would not suffer greatly. After introducing the formulation of the DBFA method,

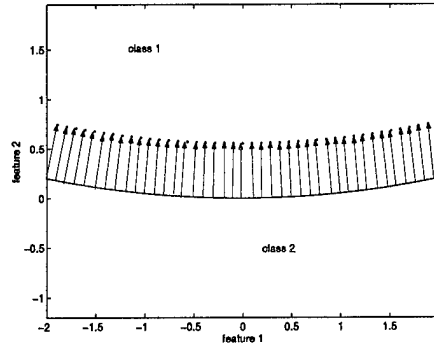


Figure 2.6 Decision boundary with unit normals

it will be used on the decision boundary of Figure 2.6 to demonstrate how this method finds the relative importance of each feature in the discrimination problem.

The basic premise behind DBFA rests upon the notion of unit normals to the decision boundary. Imagine forming unit normals (see Figure 2.6) to an arbitrary decision boundary at increments along the boundary. Now suppose the majority of the normals “point” in the same general direction. If we find some measure of the covariance of these vectors, we should be able to determine the dominant directions of the normals as well as the relative magnitudes of the other less important directions. To illustrate such a measure of covariance, consider the vectors generated in Figure 2.6. These vectors can be translated to the origin and plotted as points. This is shown in Figure 2.7. Notice we simply have a set of vectors of which we can easily find the covariance matrix from the covariance relation

$$\Omega = \frac{1}{(N-1)} \sum_i^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

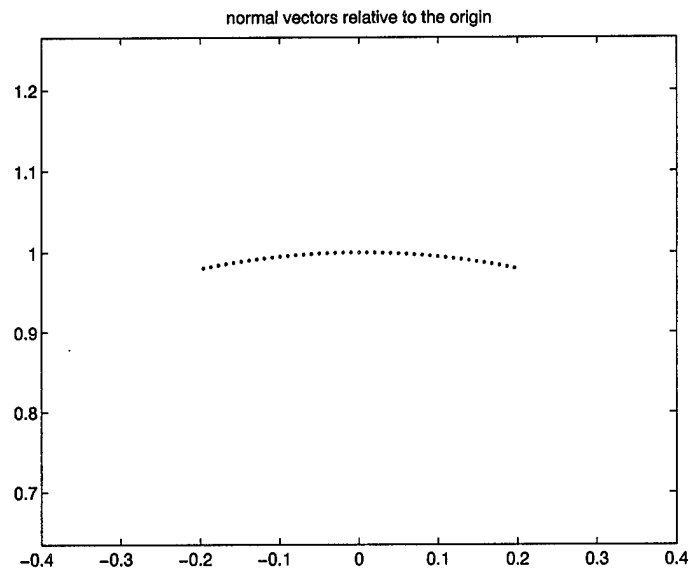


Figure 2.7 Unit normals relative to origin

where  $x_i$  and  $\bar{x}$  are the normals and their mean respectively. Calculating this covariance yields the following matrix.

$$\Omega = \begin{bmatrix} .0140 & 0 \\ 0 & 0 \end{bmatrix}$$

As before, to find the directions of maximum variance we simply find the eigenvalue and eigenvector matrices. For the above covariance matrix we have

$$\Phi = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\epsilon = \begin{bmatrix} .014 & 0 \\ 0 & 0 \end{bmatrix}$$

where  $\Phi$  and  $\epsilon$  are the eigenvector and eigenvalue matrices respectively. Immediately we see that one eigenvalue dominates and that its corresponding eigenvector is nearly horizontal. But is this what we wanted? No, we desire the direction of maximum variance to be in the direction of the “most common” normal vector. So what is the problem? Notice how the covariance was taken. This is the standard covariance relation and it always corrects for the mean. In other words, it gives a variance about the mean. Let’s suppose we form the covariance matrix without correction for the mean. The relation is simply:

$$\Omega = \frac{1}{N} \sum_i^N \mathbf{x}_i \mathbf{x}_i^T \quad (2.9)$$

Using this relation, the new (*autocorrelation* or *Effective Decision Boundary Feature Matrix* **EDBFM**) and its corresponding eigenvector and eigenvalue matrices are shown below.

$$\mathbf{EDBFM} = \begin{bmatrix} .0014 & 0 \\ 0 & .1011 \end{bmatrix}$$

$$\Phi = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\epsilon = \begin{bmatrix} .0014 & 0 \\ 0 & .1011 \end{bmatrix}$$

Observe the drastic change in the relative magnitudes of the eigenvalues as well as the complete shift in the direction of the dominant eigenvector. This eigenvector now indicates

the direction of maximum separability. The relative magnitudes of the eigenvalues together with their eigenvectors give good insight into the importance of corresponding features to classification. To clarify suppose we have a classification problem in  $\mathbf{R}^5$ . Further suppose that we have obtained a decision boundary (*discriminant function*) by some method. We then calculate unit normals at locations along the boundary and form the **EDBFM** using equation 2.9. Suppose the eigenvectors and eigenvalues of this correlation matrix are

$$\Phi = \begin{bmatrix} .2941 & .4566 & .2357 & .2063 & .6653 \\ .7352 & .1305 & .6285 & .4126 & .0739 \\ .5882 & .6523 & .3143 & .8251 & .4435 \\ .0735 & .0652 & .6285 & .3094 & .5175 \\ .1470 & .5871 & .2357 & .1031 & .2957 \end{bmatrix}$$

$$\epsilon = \begin{bmatrix} 3 & 0 & 0 & 0 & 0 \\ 0 & 8 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

if we form a weighted sum as shown below

$$\Psi = \sum_{i=1}^5 \lambda_i |\phi_i| = \begin{bmatrix} 6.4971 \\ 6.2306 \\ 11.3558 \\ 3.7547 \\ 6.3176 \end{bmatrix}$$

we see that feature 3 in the original space is the most important while features 1, 2 and 5 are roughly equivalent for discrimination. Feature 4 appears to be significantly less important for the discrimination of classes. The above heuristic has been used by Eisenbies [8] with good results. This thesis proposes an alternative heuristic (*see appendix A*) which is guaranteed to give results identical to those of Ruck's saliency calculated at the decision boundary.

## 2.6 Comments

These feature selection topics have been discussed at a very introductory level. The mathematical basis behind these transformations have been omitted but are readily available [27, 15, 22]. It may be of interest to see how the features selected using the DBFA method compare to those of other saliency metrics such as Ruck's, Tarr's or SNR. They should be identical which poses an interesting question. Which ones are easier to calculate? Tarr's may be the easiest but it requires a trained network. Although we assumed we were using a decision boundary from a trained network when we applied DBFA, this may not be

necessary. Suppose we take as the normals, the line segments connecting two close points in two different classes. This should roughly approximate the actual set of normals. This concept has been briefly studied with fair results [35].

## *2.7 Conclusion*

As was stated in the introduction, good feature selection may be the most important aspect of pattern recognition. The methods mentioned above date as far back as 1936. What has been presented here, although not nearly complete, represents the most common techniques. Although highly data dependent, it seems at present that DBFA may provide a possible measure of determining feature saliency without a trained network.

### III. Methodology

The initial problem encountered in this research stemmed from the very large dimensionality of the data set and the relatively few feature vectors. Recall from Chapter II section 2.3, Foley [14] gives a theoretical limitation to the number of features which can be used for accurate classification given a set number of exemplars. Foley's rule is restated below.

$$3(\text{number of features})(\text{number of classes}) \leq \text{number of exemplars}$$

The data set of this study contains 170 features, 2 classes and 59 exemplars, a clear violation of the Foley criterion. Given the fixed number of exemplars, Foley's rule dictates that the feature space be reduced to 10 or fewer features.

As discussed in Chapter II, there are many techniques designed to determine the most salient features from a given feature set. These were discussed in two separate categories, *classifier-based* and *classifier-free* feature saliency. Due to the gross violation of the Foley criterion, initial use of classifier-based saliency metrics was avoided. Instead, a *classifier-free* technique was employed to first reduce the feature space sufficiently to allow a non-parametric classifier to be used for further saliency analysis. This initial phase involved a classifier-free technique [21] discussed in the following section.



### 3.1 Correlation Analysis Feature Screening

Correlation Analysis is a very simple procedure for quickly determining which features (individually) in a given set are most correlated with class assignment [21]. In the application of this technique, it is hoped that the features which have the highest correlation with class will be the most salient. Determining the correlation coefficients for each feature is straightforward. Given the data is provided in a  $(p \times n)$  matrix where  $p$  is the number of exemplars and  $n$  is the number of features, simply attach an additional column which contains the classification code. Finding the correlation matrix of this  $(p \times (n + 1))$  matrix yields a new  $((n + 1) \times (n + 1))$  matrix. Extracting the first  $n$  elements of the  $(n + 1)$ st column gives the correlation coefficients for each feature.

Alternative approaches for initial screening of the 170 dimensional feature space were mentioned in Chapter II Section 2.5.1.1. In section 2.5.1.1, Fisher's discriminant ratio was introduced. This is one of the classical techniques used to screen individual features from a large feature space. Recall however, that the Fisher ratio results depend heavily on the normality of the data within each class for each dimension. In other words, it requires the data to be multivariate normal. Unfortunately, the small sample sizes within each class or possibly, the true distributions, prohibit the assumption of normality for the data of this research. As such, Fisher ratios were not used for initial screening.

Other generalized forms of the Fisher ratio were introduced in section 2.5.1.1. As before, each of these criteria were derived with the assumption of normality. Furthermore, these criteria require either the evaluation of a determinant or a matrix inverse. Due to the

extremely high correlations between many of the features, the within-class  $S_w$  and between-class  $S_b$  scatter matrices were often found to be singular. These computational difficulties prohibited the use of these techniques for initial screening. The above considerations precipitated the use of correlation analysis for initial screening.

### *3.2 Cascade-Correlation on the Input Features*

It was mentioned in section 3.1, that many of the features from this database are highly correlated. This is a product of the extraction method used for these features. For instance, features 1-25 are all Angular Second Moment features. The extraction procedure for each is very similar. The result is that these features are highly correlated with each other. In fact, features of the same *type* such as Angular Second Moment, have correlations near 1. With this in mind, a procedure was developed which would identify those features which are highly correlated with class while having minimal correlation between them. The algorithm is straightforward and easy to implement in MatLab. The following pseudo-code illustrates the approach.

1. Find the feature  $X_c$  with the largest correlation with class
2. Eliminate all features which are correlated with  $X_c$  above a user specified threshold  $C_t$
3. While features remain Return to step 1
4. End

At the conclusion of this sequence, a set of features is obtained which have pairwise correlations below the threshold. The procedure is then repeated with slightly larger thresholds until a sufficient number of features have been identified. The features identified at lower thresholds are very nearly independent but not necessarily good discriminators. Those obtained at high thresholds are good discriminators but are likely to be highly correlated. The theoretical premise is that the procedure will give a good pool of features in which to apply more sophisticated saliency methods. For this research, the cascade correlation procedure was used to identify 20 potentially salient features from the initial 170.

### 3.3 *Classifier-Based Saliency*

Section 2.5.2.1 introduced four classifier-based saliency techniques. This research implemented three of these as well as two additional techniques to be described below. The three previously discussed techniques are found in section 2.5.2.1.

**3.3.1 *Ruck Saliency at Data Points.*** Calculation of Ruck's saliency as proposed in section 2.5.2.1, requires the evaluation of partial derivatives at a large number of points in the input space. This number grows rapidly with the dimension of the feature space

and as such, has prompted many to seek effective application of Ruck's saliency with fewer calculations. One such approach requires the evaluation of Ruck's saliency only where there is actual data [34]. This is expressed below.

$$\Lambda_i = \sum_p \sum_k \left| \frac{\partial z_k}{\partial x_i} \right| \quad (3.1)$$

here,  $k$  is the index over all outputs and  $p$  is the index over all feature vectors in the training set. Again the absolute value of the partial derivative is used because we are only concerned with the magnitude of the change in outputs. The sum over all outputs is necessary to measure the full sensitivity to a change in input. Here, the partial derivatives are evaluated only at known data points in the feature space. With  $\Lambda_i^n$  representing the saliency of feature  $i$  calculated from the  $n$ th MLP, the average over  $N$  MLP's is given by

$$\Lambda_i = \frac{1}{N} \sum_{n=1}^N \Lambda_i^n \quad (3.2)$$

**3.3.2 Ruck Saliency at the Decision Boundary.** Another variant of Ruck's saliency to be introduced in this research involves the decision boundary directly. This method involves first finding a set of vectors  $\mathbf{P}$  which lie on the decision boundary or near the decision boundary within a given tolerance. Ruck's saliency is then calculated using this set  $\mathbf{P}$  in equation 3.1. The following pseudo-code illustrates the process.

1. Train neural network

2. Remove all exemplars which are misclassified
3. For every vector in class 1, find it's nearest neighbor in class 2
4. Form a line connecting these two points
5. Find the point  $p$  along the line where classification changes
6. Repeat step 3 for class 2

The advantage of this method is that it only requires partial derivatives to be computed at  $P$  points where  $P$  is the number of exemplars. This is the same number of evaluations as Ruck at the data points, however, the set of points  $P$  must first be calculated. In practice, this has not been difficult or computationally expensive. Comparisons between performance of each of the Ruck variants will be given as part of this research.

### 3.4 Summary

This chapter introduced the *correlation-with-class* concept as a feature saliency technique. In addition, the technique was extended in an effort to acquire *pairwise* uncorrelated features using the cascade algorithm. Two classifier-based techniques, Ruck's at the known data and Ruck's at the decision boundary were also introduced. These methods will be used in combination in an attempt to optimize the feature set.

## IV. Analysis and Results

### 4.1 Database

The database consists of 59 observations for which we have truth data. These fall into two classes, malignant (class 1) and benign (class 2). There are 23 observations in class 1 and 36 observations in class 2. Each observation is a feature vector ( $1 \times 170$ ). The 170 features which have been extracted from each ROI fall into the categories as shown in Table 4.1.

Col	Feature
1-25	Angular Second Moment
26-49	2nd Order Contrast
50-74	2nd Order Entropy
75-99	Correlation
100	Average Distance Between Calcifications
101	Standard Deviation of Gray Levels
102	Mean of Gray Levels
103	1st Order Contrast
104	1st Order Entropy
105-129	Laws Energy Ratios
130-135	Power Spectrum Rings
136-155	Wavelet Coefficients
156-170	Eigenmass Coefficients

Table 4.1 Features by type

### 4.2 Initial Screening

The analysis of data began with the initial screening of the 170 features using correlation analysis. Computation of the correlations is straightforward. The results are shown in Figure 4.1. Only the magnitude is of concern, therefore the results shown are the correlation

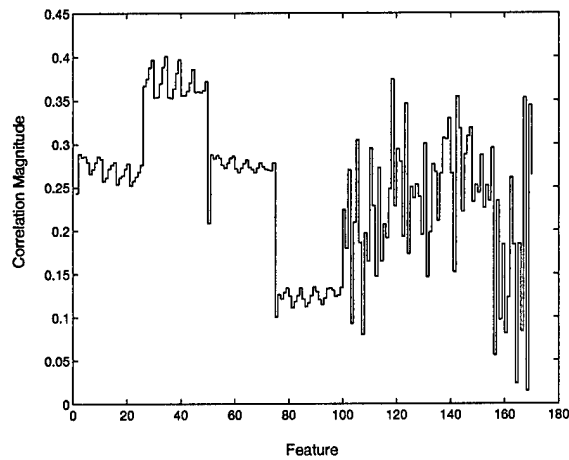


Figure 4.1 Correlations of features with class

magnitudes.

In addition to the calculation of correlations, Fisher's ratio was calculated for comparative purposes. The Fisher discriminant ratio and the correlation screening techniques provided similar sets of top 20 features. Table 4.2 gives the top twenty (*in descending order*) from each method.

Correlation	Fisher
123	123
127	127
29	29
34	34
28	100
39	28
33	27
27	26
44	39
38	33
129	129
49	102
26	30
30	170
43	44
32	38
35	140
48	136
45	32
40	35

Table 4.2 Top 20 features from initial screening

### 4.3 Initial Classification

Selecting the top ten features from correlation analysis to use in an MLP classifier produced acceptable results. A neural network with 3 hidden layer nodes and two output nodes was used. Sigmoid activation functions were used on the hidden layer and output nodes. MatLab's accelerated backpropagation (**trainbpx**) was the training algorithm of choice. The overall classification accuracy using the hold-one-out method was 74.6% obtained as an average over 10 realizations of the hold-one-out method. Table 4.3 summarizes the results. A **False -** is a malignant exemplar classified as benign while a **False +** is a benign exemplar classified as malignant. For comparative purposes, the top 10 features from the

Avg False -	Avg False +	Avg Accuracy %
5.2	9.8	74.6

Table 4.3 Classification results using top 10 correlation features

Fisher ranking were used in the same network. Although the two sets differed by only two features, there was a substantial difference in the error rate. Table 4.4 summarizes the results.



Avg False -	Avg False +	Avg Accuracy %
7.4	13.6	64.4

Table 4.4 Classification results using top 10 Fisher features

For comparative purposes 10 features were selected pseudo-randomly. Every 17th feature was chosen producing the pseudo-random set of ten. Using the same error estimation as before, the following results were produced.

Avg False -	Avg False +	Avg Accuracy %
10.4	16.4	54.57

Table 4.5 Classification results using ten random features

*4.3.0.1 Examination of Correlations.* It was noted that most of the top ten features from the correlation analysis came from the same *type* as given in Table 4.1. It therefore seemed likely that these ten would be correlated, possibly providing redundant information. To check this, the correlation matrix of the features was examined. The matrix is very large( $170 \times 170$ ) and therefore hard to examine. Utilizing the visualization capabilities of MatLab, a pseudo-color map for the matrix was generated. This allowed the entire matrix to be viewed at a glance and is shown in Figure 4.2. Lighter regions correspond to high correlations.

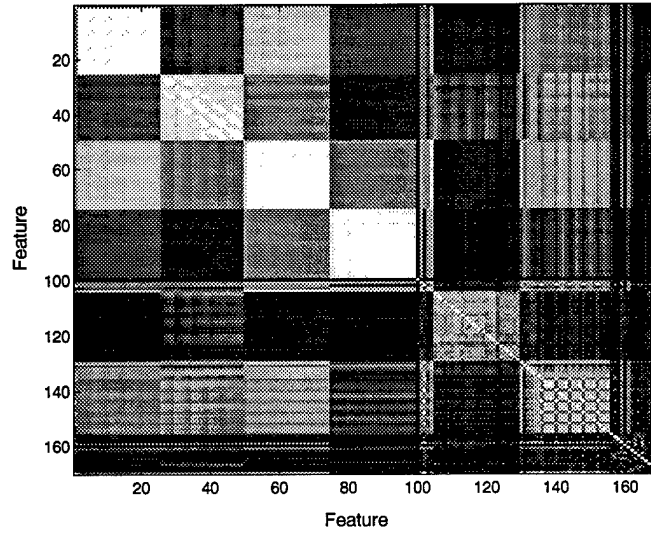


Figure 4.2 Correlation matrix

Immediately one can see that features of the same type are almost perfectly correlated.

*4.3.1 Top 20 Uncorrelated Features.* Under the assumption that highly correlated features provide redundant information (*see appendix E*), an effort was made to select features which have high correlation with class but small pairwise correlation. To this end, the cascade correlation approach as outlined in Chapter III, section 3.2 was applied. The resulting top 20 features from this procedure are given in Table 4.6.  $C_t$  is the correlation threshold sequence.

Feature	C <sub>t</sub>
34	.25
169	
113	
157	
100	
131	
158	
118	.30
165	.40
166	
164	.45
102	
163	
156	.50
101	.55
167	.60
162	
104	
123	.65
2	

Table 4.6 Pool of 20 features using cascade correlation

#### 4.4 Classifier-Based Feature Saliency

Reduction of the feature set in Table 4.6 to 10 or fewer features was accomplished through the use of the classifier-based saliency techniques as outlined in Chapters II and III. Figure 4.3 illustrates the filtering process used to identify the 10 most salient features. The neural network architecture consisted of 3 hidden layer nodes, 2 output nodes and 20 input nodes. Sigmoid activation functions were used on the hidden layer nodes and the output nodes. The network was trained 50 times using the entire data set with presentation order being randomized. Each of the 5 saliency metrics as well as the confusion matrix were calculated for each trained network. The average saliency metrics were then used to determine the top 10 features selected by each metric. The process was repeated for *each* of these sets of 10. The top six from each of these were retained. From the five sets of six, the frequency of occurrence of each feature in the five sets of six was used to rank the features

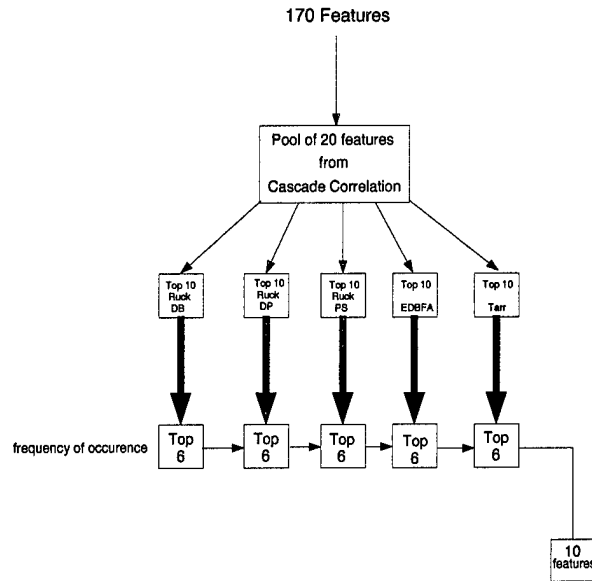


Figure 4.3 Saliency flow chart

contained within the five sets. The five sets of six constituted exactly ten features. The ten features are shown in Table 4.7 with their frequency of occurrence. Given these ten features,

Feature	frequency
163	5
34	5
104	4
165	4
102	4
101	3
123	2
156	1
100	1
131	1

Table 4.7 Top 10 features by frequency of occurrence

classification accuracy improved significantly. The results are shown in Table 4.8. Not only

Avg False -	Avg False +	Avg Accuracy %
3.75	5.75	83.9

Table 4.8 Classification results using top 10 features from combined techniques

has the overall classification accuracy improved, but both the probability of a false negative

and false positive have decreased. The classification is well balanced as well. Classification for both malignant and benign regions is approximately the overall average of 84%.

#### *4.5 Comparative Analysis of Ruck Variants*

The previous approach proved very fruitful but questions remain as to whether the initial screening techniques are necessary. Or the real question may be whether or not one could simply violate Foley's rule by training a neural network with all 170 features, then using classifier-based saliency techniques, identify a candidate subset of salient features to use in more fine tuned reduction. To answer this question, a neural network was trained 30 times over the entire data set. The network was of the same architecture as in all previous analyses only it now has 170 input nodes. The data point and decision boundary variants of Ruck's saliency were calculated for each network and averaged over the 30 trials. Ruck's saliency using pseudo-sampling was entirely impractical for this problem. To illustrate, with 59 exemplars, 170 dimensions and a modest sampling interval of 20, 34.1 million partial derivatives must be evaluated if pseudo-sampling is to be used. This involves not only the calculation of the derivatives, but the evaluation of the neural network at over 200,000 vectors. Initial runs required over 15 minutes for each network on an ULTRASPARC. This is a perfect example of why previous research has proposed that Ruck's saliency be calculated only at the data. The results of this approach are provided in Table 4.9.

Decision Boundary	Data Points
127	127
163	163
100	109
164	100
109	164
123	131
133	123
169	133
157	120
131	124

Table 4.9 Top 10 features using two variants of Ruck's saliency

Again, to test the performance of these two sets, the architecture from previous analyses was used. The results for the set obtained from Ruck's saliency at the data points are in Table 4.10. The 10 features obtained from Ruck's saliency at the decision boundary

Avg False -	Avg False +	Avg Accuracy %
9.5	11.17	<b>64.97</b>

Table 4.10 Classification results using top 10 features from Ruck's at data points

produced the results in Table 4.11. The results of each of these methods are inferior to the

Avg False -	Avg False +	Avg Accuracy %
8.17	8.17	<b>72.33</b>

Table 4.11 Classification results using top 10 features from Ruck's at decision boundary

results of section 4.4. In fact, these results are worse than the results using the top 10 *most correlated with class* features and the latter didn't require a trained classifier.

#### 4.6 Summary

This chapter has demonstrated the viability of correlation analysis as an initial screening technique when one is confronted with a high dimensional feature space. The results

of this classifier-free technique were shown to be superior to derivative-based techniques by 2 percentage points. In addition, classification accuracy was further improved using correlation analysis *followed* by classifier-based saliency techniques. This combined technique produced classification accuracies 12 percentage points better than the best derivative-based technique. Also, comparisons between two variants of Ruck's saliency were presented. It was demonstrated that Ruck's saliency calculated at the decision boundary yielded a feature set significantly better for classification than the feature set obtained using Ruck's saliency calculated at the known data. This was quantified by an increase in classification accuracy of nearly 7.5 percentage points.

## *V. Conclusions*

The goal of this thesis was to advance correlation analysis as proposed by Greene [21], and to determine its viability as an effective classifier-free initial screening technique. It has been shown for this data set, that correlation analysis is beneficial and possibly the best approach. Although these results could be highly data dependent, it is believed that the correlation analysis is robust. A summary of the key obstacles and corresponding conclusions of this thesis is given below.

### *5.1 Violation of Foley's Rule*

The small sample size as well as computational expense dictated the reduction of the feature space dimensionality. Classifier based saliency techniques depend on a well trained network for accurate results. A network trained in violation of Foley's rule will give a large variance in the computed decision boundary and response surface making the use of classifier-based saliency as a first cut very suspect. This is the motivation for a classifier-free initial screening technique.

### *5.2 Fisher's Discriminant Ratio*

One of the most well known and widely used classifier-free saliency techniques is the Fisher ratio and its generalized forms. It was previously stated that this metric assumes multivariate normality of the data. The data of this research was not multivariate normal. In fact, many of the marginal distributions were highly asymmetric. Furthermore, the gener-



alized variants require the calculation of matrix inverses or determinants. For this data, the scatter matrices were singular due to the high correlations of the features. For comparative purposes, the individual Fisher ratios were calculated and the top 10 used for classification. It was demonstrated that these features were significantly worse than those obtained from correlation analysis.

### *5.3 Correlation Analysis*

Correlation analysis is straightforward and computationally inexpensive. In addition, it makes no assumptions about the data. One drawback is that it provides only linear correlations while eliminating features which may have some useful nonlinear correlation with class. Also, it may select features which are highly correlated with themselves, possibly providing redundant information. Nevertheless, the results proved significantly better (*10 percentage points*) than the results using features obtained from the Fisher ranking.

### *5.4 Cascade Correlation Analysis*

To eliminate the retention of correlated features in the retained set, a new "cascade" correlation method was developed and implemented. This was combined with classifier-based saliency to produce a very salient set of ten features. The results from this combined method were significantly better than the results from any other method alone. The highest classification accuracy from any stand alone method was 74.6% while the combined method produced an average classification accuracy of 83.9%.

### 5.5 *Known Data vs. Decision Boundary*

Two variants of derivative-based saliency were employed. Classification results indicate that derivative-based saliency at the decision boundary is superior to derivative-based saliency at the known data. This likely stems from the high variability of the response surface at locations far from the decision boundary (*see appendix D*). Although derivative-based saliency at the decision boundary requires calculation of the decision boundary, it is computationally trivial, and vastly more practical than pseudo-sampling when the dimensionality of the feature space is high.

## *Appendix A. Ruck at the Decision Boundary vs. EDBFA*

EDBFA as introduced by Lee and Landgrebe [22] proposed no procedure for rank ordering a feature set. This has remained an open question. Heuristics such as the one in chapter II, section 2.5.2.1 have been adopted which produce good results but which may not be an optimal use of the information provided by the eigenvector/eigenvalue pairs. This thesis proposes that there is a way to combine the eigenvalue/eigenvector pairs which should produce identical results with those of Ruck's saliency at the decision boundary. Figure A.1 illustrates a simple two-class problem in two dimensions. In this figure,  $\circ$  represents an

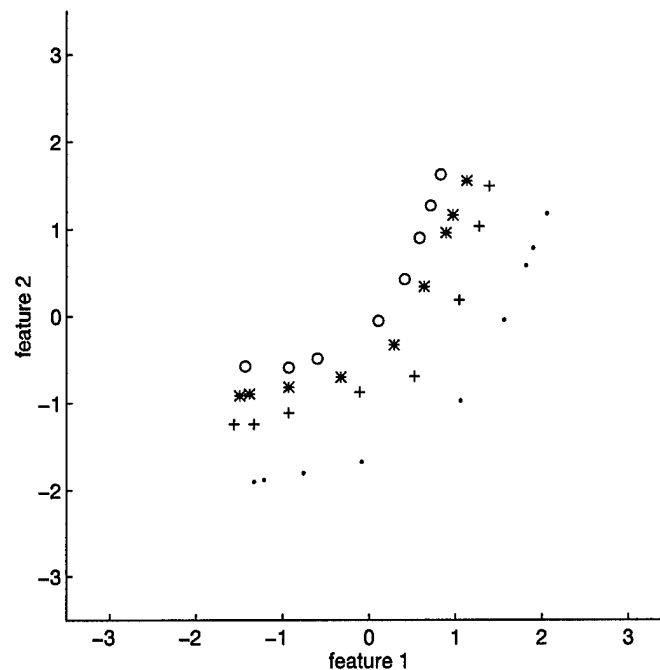


Figure A.1 Decision boundary for two dimensional two class problem

exemplar from class 1 and + is an exemplar from class 2. The \*'s are the calculated decision

boundary. The  $\cdot$ 's are the endpoints of the normal vectors originating at each of the decision boundary points. The ray itself is not drawn. Notice, intuitively we would expect each of the features to be important for discrimination. The normals at each of the decision boundary points are given in Table A.1. Ruck's saliency at the decision boundary is simply the sum

$\frac{\partial z}{\partial x_1}$	$\frac{\partial z}{\partial x_2}$
0.9240	-0.3823
0.9260	-0.3774
0.2412	-0.9705
0.9261	-0.3774
0.9261	-0.3774
0.1688	-0.9857
0.7685	-0.6399
0.1700	-0.9854
0.7685	-0.6399
0.1700	-0.9854
0.9261	-0.3774
0.1687	-0.9857
0.1688	-0.9857
0.9261	-0.3774
0.2412	-0.9705
0.9240	-0.3823

Table A.1 Normal vectors

of the absolute magnitudes of each column. Consider the possibility of summing the squares rather than the magnitudes. This should give different numerical values but the saliency *ranking* should stay the same. These results are in Table A.2.

$\Sigma \frac{\partial z}{\partial x_1}$	$\Sigma \frac{\partial z}{\partial x_2}$	$\Sigma (\frac{\partial z}{\partial x_1})^2$	$\Sigma (\frac{\partial z}{\partial x_2})^2$
9.3439	10.8002	<b>7.4362</b>	<b>8.5638</b>

Table A.2 Ruck at the decision boundary

This indicates that the two features are approximately equal in saliency. To implement the EDBFA approach, data from Table A.1 is put into a matrix  $\mathbf{X}$ , and the EDBFM is calculated as

$$\mathbf{EDBFM} = \mathbf{X}^T \mathbf{X}$$

which yields

$$\mathbf{EDBFM} = \begin{bmatrix} \mathbf{7.4362} & -4.7397 \\ -4.7397 & \mathbf{8.5638} \end{bmatrix}$$

Notice the diagonal entries are just the sums of squares as found in Table A.2. They are of course, the variances (*non-centered*) of the components of the normals. This thesis questions whether proceeding with Lee and Landgrebe [22] decision boundary analysis is even necessary. Isn't all the information we need contained in these diagonal elements? These diagonals provide the same information as Ruck's saliency. The only additional information from this matrix is the covariance. But are the covariances important? They do affect the eigenvalues and eigenvectors but Lee and Landgrebe [22] never proposed how one should interpret these eigenpairs in terms of feature saliency. The fact is, the total variance does not change. The eigenvalues are the variances along the eigenvectors. The magnitudes of the components of the eigenvectors indicate how much of the variance is in that components

dimension. But this information is already summed up in the diagonal elements. This argument suggests that DBFA provides no additional information from that already available in the Ruck analysis.

Next, a heuristic will be proposed for ranking features in DBFA. It will be shown that this method yields **exactly** the same results as Ruck's saliency calculated at the decision boundary. In section 2.5.2.1, a heuristic was proposed for feature ranking using DBFA. The heuristic involved an eigenvalue-weighted sum of the eigenvectors and is reproduced below.

$$\Psi = \sum_i \lambda_i |\phi_i|$$

Suppose the absolute value operator is replaced with the "squared" operator  $(\cdot)^2$  such that every element of the vector  $\phi$  is squared. This is shown below and has been proposed by Stewart [35].

$$\Psi = \sum_i \lambda_i (\phi_i)^2 \tag{A.1}$$

This method will guarantee the same results as given in Table A.2. To prove this, one only has to look at the spectral decomposition of the matrix **EDBFM**. The decomposition is formed from the eigenvalue-weighted sum of the outer products of the eigenvectors [7]. Each of the outer products forms a matrix with diagonal elements equal to the squares of the elements of the generating eigenvector. The resulting sum of these matrices will of course yield **EDBFM**. Therefore, if one were to square the elements of each eigenvector prior to forming the weighted sum, the results would be precisely the diagonals of **EDBFM**.

(see *appendix B*). But these diagonals are exactly the sums of squares found in Table A.2. Furthermore, these sums of squares have the same ranking as the sums of absolute values as originally proposed by Ruck. This guarantees the equivalence of Ruck's saliency calculated at the decision boundary and DBFA saliency using the heuristic in equation A.1.

## *Appendix B. Spectral Decomposition*

### *Proof 1:*

Suppose Ruck's saliency has been calculated at  $n$  points along the decision boundary. Each of these  $n$  gradients can be normalized and placed in a matrix  $\mathbf{X}$  such that each row is a normal vector and each column is the components partial derivative. From this matrix, Ruck's saliency is calculated from the sums of the absolute values of each column. For notational simplicity, take the two-dimensional case as shown below.

$$\mathbf{X} = \begin{matrix} & \frac{\partial z}{\partial x_1} & \frac{\partial z}{\partial x_2} \\ \begin{pmatrix} a_{11} & a_{12} \\ \vdots & \vdots \\ a_{n1} & a_{n2} \end{pmatrix} \end{matrix}$$

Ruck's saliency is then,

$$\Lambda_1 = \sum_N \left| \frac{\partial z}{\partial x_1} \right| \qquad \Lambda_2 = \sum_N \left| \frac{\partial z}{\partial x_2} \right| \qquad (\text{B.1})$$



or,

$$\Psi = \begin{bmatrix} \Lambda_1 \\ \Lambda_2 \end{bmatrix}$$

The ranking of the components of  $\Psi$  is the Ruck ranking of features.

Suppose the sums of the squares of the columns had been used rather than the sums of the absolute values. The Ruck saliency of each feature is now,

$$\Lambda_1 = \sum_N \left( \frac{\partial z}{\partial x_1} \right)^2 \quad \Lambda_2 = \sum_N \left( \frac{\partial z}{\partial x_2} \right)^2 \quad (\text{B.2})$$

The EDBFM as proposed by Lee and Landgrede [22] is defined to be the outer product of the matrix  $\mathbf{X}$ . Calling this matrix **EDBFM** gives the following relation,

$$\mathbf{EDBFM} = \mathbf{X}^T \mathbf{X} = \begin{bmatrix} A & B \\ B & D \end{bmatrix}$$

which is symmetric. Note also that the diagonal elements  $A$  and  $D$  of **EDBFM** are the sums of squares of the columns of  $\mathbf{X}$  so that,

$$\mathbf{EDBFM} = \begin{bmatrix} A & B \\ B & D \end{bmatrix} = \begin{bmatrix} \Lambda_1 & B \\ B & \Lambda_2 \end{bmatrix}$$

and,

$$\Psi = \begin{bmatrix} \Lambda_1 \\ \Lambda_2 \end{bmatrix} = \begin{bmatrix} A \\ D \end{bmatrix} \quad (\text{B.3})$$

The spectral decomposition of a symmetric matrix guarantees that the matrix **EDBFM** can be decomposed into a linear combination of the outer products of its orthonormal eigenvectors [7]. If  $v_1$  and  $v_2$  are the orthonormal eigenvectors of **EDBFM** and  $\lambda_1$  and  $\lambda_2$  are the corresponding eigenvalues, the matrix **EDBFM** can be decomposed as shown,

$$\mathbf{EDBFM} = \begin{bmatrix} A & B \\ B & D \end{bmatrix} = \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T$$

Now, if  $v_1$  and  $v_2$  are written in component form,

$$v_1 = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \quad v_2 = \begin{pmatrix} \gamma \\ \delta \end{pmatrix}$$

and the spectral decomposition expanded,

$$\mathbf{EDBFM} = \begin{bmatrix} A & B \\ B & D \end{bmatrix} = \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T = \lambda_1 \begin{bmatrix} \alpha^2 & \alpha\beta \\ \beta\alpha & \beta^2 \end{bmatrix} + \lambda_2 \begin{bmatrix} \gamma^2 & \gamma\delta \\ \delta\gamma & \delta^2 \end{bmatrix}$$

so that,

$$A = \lambda_1 \alpha^2 + \lambda_2 \gamma^2$$

$$D = \lambda_1 \beta^2 + \lambda_2 \delta^2$$

or,

$$\begin{aligned} \begin{bmatrix} A \\ D \end{bmatrix} &= \lambda_1 \begin{pmatrix} \alpha^2 \\ \beta^2 \end{pmatrix} + \lambda_2 \begin{pmatrix} \gamma^2 \\ \delta^2 \end{pmatrix} \\ &= \lambda_1 (v_1)^2 + \lambda_2 (v_2)^2 \\ &= \sum_I \lambda_i (v_i)^2 \end{aligned} \tag{B.4}$$

and so by equations B.3 and B.4,

$$\Psi = \begin{bmatrix} \Lambda_1 \\ \Lambda_2 \end{bmatrix} = \sum_I \lambda_i (v_i)^2 \tag{B.5}$$

This demonstrates the equivalency of the Lee and Landgrebe [22] approach using the metric of equation B.5 and the Ruck saliency as proposed in equation B.2. All that remains to be shown is that the Ruck ranking of features using the absolute value as in equation B.1 does not change if the ranking is calculated using the sums of squares as in equation B.2. This proof [24] is provided in appendix C.

### *Appendix C. Ranking Equivalency*

#### *Proof 2:*

Given a matrix of elements  $x$ , such that  $|x| \leq 1$ ,

$$\begin{bmatrix} a & c \\ b & d \end{bmatrix}$$

constrained by the normalization relations,

$$a^2 + c^2 = 1 \qquad b^2 + d^2 = 1 \qquad (\text{C.1})$$

show that,

$$\text{if } |a| + |b| \geq |c| + |d| \quad \text{then} \quad a^2 + b^2 \geq c^2 + d^2 \qquad (\text{C.2})$$

Begin by incorporating the constraint relations in equations C.1 into the matrix.

$$\begin{bmatrix} a & \sqrt{1-a^2} \\ b & \sqrt{1-b^2} \end{bmatrix}$$

Now rewrite the hypothesis C.2,

$$\text{if } |a| + |b| \geq \sqrt{1-a^2} + \sqrt{1-b^2} \text{ then } a^2 + b^2 \geq 2 - a^2 - b^2 \quad (\text{C.3})$$

Assume C.3 is not true such that,

$$a^2 + b^2 < 2 - a^2 - b^2 \quad (\text{C.4})$$

then

$$a^2 + b^2 < 1 \quad (\text{C.5})$$

Now since  $\overbrace{|a| + |b|}^x \geq 0$  and  $\overbrace{\sqrt{1-a^2} + \sqrt{1-b^2}}^y \geq 0$ , then  $x^2 \geq y^2$ , and the following inequality can be written,

$$a^2 + |ab| + b^2 \geq 1 - a^2 + 1 - b^2 + \sqrt{(1-a^2)(1-b^2)} \quad (\text{C.6})$$

$$a^2 + |ab| + b^2 \geq 2 - a^2 - b^2 + \sqrt{1 - a^2 - b^2 + a^2b^2} \quad (\text{C.7})$$

$$2(a^2 + b^2) + |ab| \geq 2 + \sqrt{1 - a^2 - b^2 + a^2b^2} \quad (\text{C.8})$$

now by C.5, it is true that,

$$2(a^2 + b^2) < 2 \quad (\text{C.9})$$

combining equations C.8 and C.9 gives the relation,

$$|ab| > \sqrt{1 - (a^2 + b^2) + a^2b^2} \quad (\text{C.10})$$

again by condition C.5, it must be true that,

$$\sqrt{1 - (a^2 + b^2) + a^2b^2} > \sqrt{a^2b^2} \quad (\text{C.11})$$

Combining relations C.10 and C.11 gives the following false relation,

$$|ab| > \sqrt{a^2b^2} \quad (\text{C.12})$$

This implies that relation C.5 cannot hold, which proves that inequality C.4 is false when the left side of inequality C.3 is true. Therefore the hypothesis C.2 is proved. It is a simple matter to show this proof holds if columns are added to the matrix. It can also be shown by induction that the proof will hold if rows are added.

### *Appendix D. Ruck at the Decision Boundary vs. Ruck at Known Data*

The results of this thesis are insufficient to claim that Ruck's saliency at the decision boundary will always outperform Ruck's saliency at the known data. However, it does suggest that it will likely outperform the *known data* variant if the network is poorly trained, while giving equivalent results otherwise. An example of such possibilities is given next.

Figure D.1 shows a two dimensional two-class problem with the accompanying decision boundary. The problem is linearly separable and should only require a single hidden layer node. To illustrate the implications of improper architecture or improper training, a network

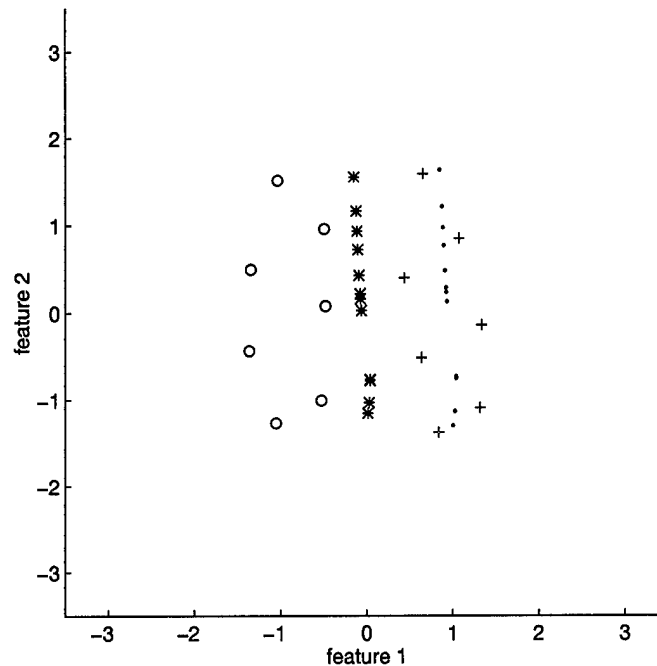


Figure D.1 Two dimensional two-class problem

with 7 hidden layer nodes was used. This produces more variation in the resulting response

surface since the network attempts to use all of its nodes. Figure D.2 demonstrates this “noisy” surface. The original data is shown on the lower plane while the zero plane cuts

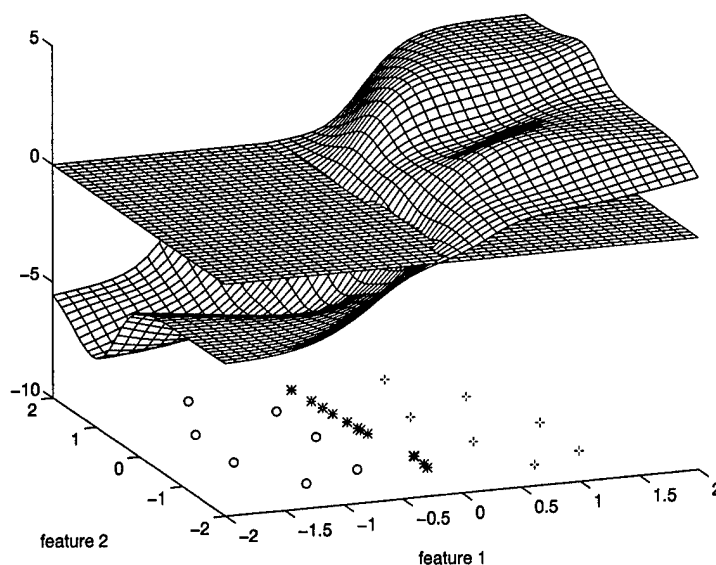


Figure D.2 Neural network response surface

the surface at the decision boundary. Notice (*by inspection*) that the partials taken at the data points have significant magnitudes in both dimensions. At the decision boundary however, there is much less variation. The significant component of the gradients is much less ambiguous. To quantify these conclusions, the saliency for each method was computed for this problem. The results are given in Table D.1. Although the two methods properly

Method	feature 1	feature 2	Ratio
DB	39.8922	6.7481	<b>5.9116</b>
DP	3.3610	1.7255	<b>1.9478</b>

Table D.1 Ruck saliency values

identify the most salient feature, Ruck's at the data points is much less certain about which



is better. In a two dimensional problem, it is unlikely that it would rank these erroneously but in a 170 dimensional problem, it is highly likely that the ranking will differ. This is in fact what was seen for the data used in this thesis.

To demonstrate the convergence of the two techniques when the neural network is properly trained, another example is given. In this example, the Ruck saliency using pseudo-sampling is calculated as well. Figure D.3 illustrates the two dimensional two-class problem. Again, the problem is clearly linearly separable. Proper architecture requires only one hidden

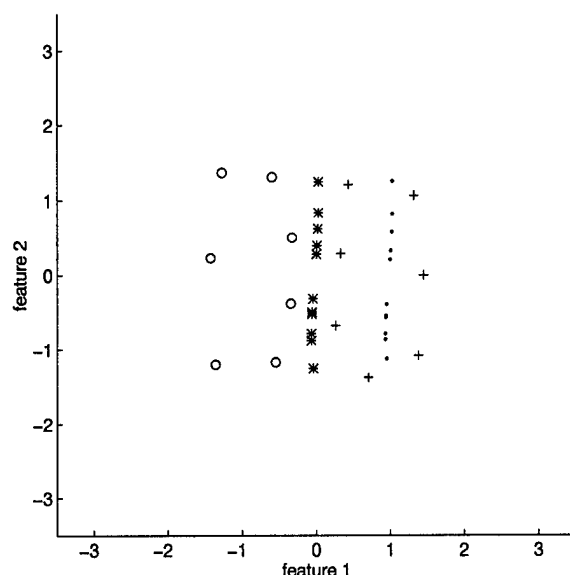


Figure D.3 Two class pattern distributions

layer node. Additional nodes provide too many degrees of freedom. The neural network output using 7 nodes and 1 node respectively is shown in Figures D.4 and D.5. The saliency results are given in Tables D.2 and D.3.

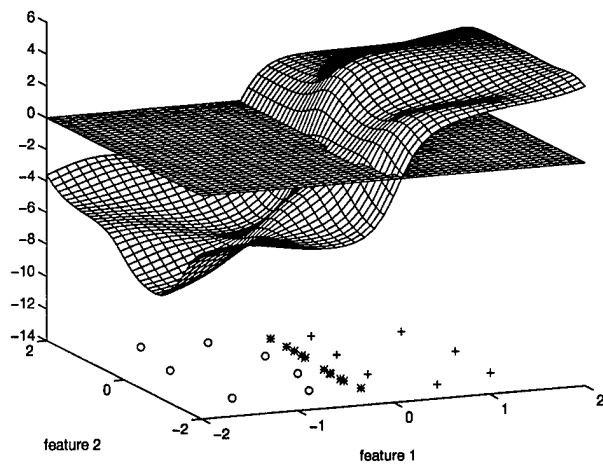


Figure D.4 Poorly trained network

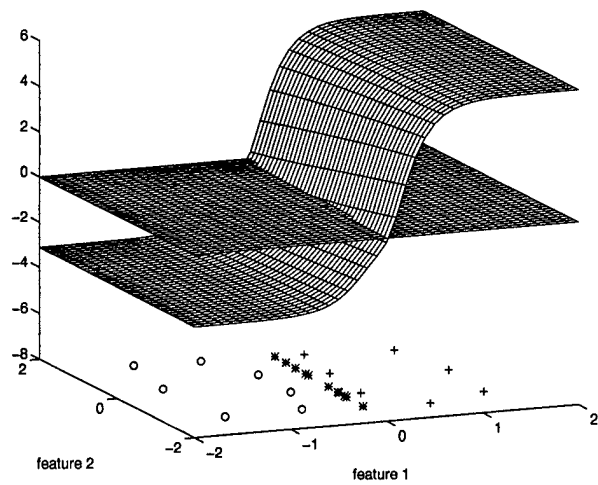


Figure D.5 Properly trained network

Method	feature 1	feature 2	Ratio
DB	68.3818	5.3170	<b>12.8610</b>
DP	3.5258	1.5659	<b>2.2516</b>
PS	385.4236	99.7109	<b>3.8654</b>

Table D.2 Ruck saliency with poorly trained network

Method	feature 1	feature 2	Ratio
DB	84.5169	3.1065	<b>27.2065</b>
DP	3.9458	.1450	<b>27.2124</b>
PS	399.9759	14.7018	<b>27.2059</b>

Table D.3 Ruck saliency with properly trained network

The ratios indicate that the saliency metrics converge when the network is properly trained. It is again evident that Ruck's saliency calculated at the decision boundary suggests a far greater importance of feature 1, relative to feature 2, than Ruck's saliency using the other variants. In other words, Ruck's saliency at the decision boundary when the network is improperly trained, gives a ratio more consistent with that obtained from the saliency ratios of a properly trained network.

## *Appendix E. Should Cascade Correlation Work?*

One of the fundamental premises behind the cascade correlation concept is that features which are highly correlated with each other provide redundant information, allowing one to be discarded. Figure E.1 illustrates a two dimensional two-class problem in which the features are highly correlated but both are necessary for accurate classification. Feature 1 and feature

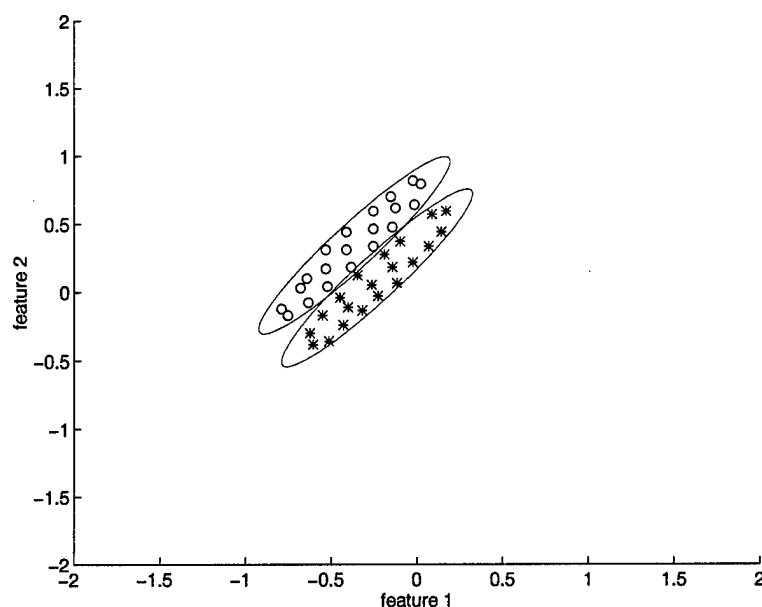


Figure E.1 Highly correlated features

2 have correlations with class of .2606 and -.4309 respectively. The correlation between them is .7248. In this case, cascade correlation may select feature 2 because of its high correlation with class, then discard feature 1 because it is highly correlated with feature 2. This would cause a severe reduction in classification accuracy. Whether or not the procedure will discard

feature 1 depends entirely on the sequence of values  $C_t$  used in the procedure. These values are chosen arbitrarily however. It seems entirely possible that a better use of correlation information can be found and exploited.

## Bibliography

1. Bellman, R.E. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961.
2. Belue, L.M. and K.W. Bauer. "Determining Input Features for Multilayer Perceptrons," *Neur. Comput.*, 7 no. 2 (1995).
3. Chandrasekaran, B. and A.K. Jain. *Handbook of Statistics, Vol. 2*. North Holland Publishing Company, 1982.
4. Chang, T. and C.J. Kuo. "Texture Analysis and Classification with Tree Structured Wavelet Transform," *IEEE Transactions on Image Processing*, 2 (October 1993).
5. C.M. Kocur, S.K. Rogers, K.W. Bauer et al. "Using Neural Networks to Select Wavelet Features for Breast Cancer Diagnosis," *IEEE Engineering in Medicine and Biology* (May/June 1996).
6. Dauk, Capt. Ronald C. *Computer Aided detection of Microcalcifications Using Texture Analysis*. MS thesis, Air Force Institute of Technology, OH, December 1995.
7. Dillon, W.R. and M. Goldstein. *Multivariate Analysis: Methods and Applications*. John Wiley and Sons, 1984.
8. Eisenbies, Christopher Lawrence. *Classification of Ultra High Range Resolution Radar Using Decision Boundary Analysis*. MS thesis, Air Force Institute of Technology, OH, December 1994.
9. et al., A.F. Laine. "Mammographic Feature Enhancement by Multiscale Analysis," *IEEE Transactions on Image Processing*, 13 (December 1994).
10. et al., A.P. Dhawan. "Artificial Neural Network Based Classification of Mammographic Microcalcifications Using Image Structure Features," *SPIE*, 1905 (1993).
11. et al., D.W. Ruck. "Feature Selection Using a Multilayer Perceptron," *Journal of Neural Network Computing*, 2(2) (1990).
12. et al., Yuzheng Wu. "Artificial Neural Networks in Mammography: Application to Decision Making in the Diagnosis of Breast Cancer," *Radiology*, 187 (April 1993).
13. Fisher, R.A. "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 7 (1936).
14. Foley, D.H. "Considerations of Sample and Feature Size," *IEEE Transactions on Information Theory*, IT-18 (1972).
15. Fukunaga, Keinosuke. *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.
16. Giger, Maryellen. "Computer-Aided Diagnosis." RSNA Categorical Course in Physics, 1993.

17. Goldstein, U. "Speaker Identifying Features Based on Formant tracks," *JASA*, 59(1) (January 1976).
18. H. Yoshida and R.M. Nishikawa. "Automated Detection of Clustered Microcalcifications in Digital Mammograms Using Wavelet Transform Techniques," *SPIE Image Processing*, 2167 (1994).
19. K.A. Greene, K.W. Bauer, D.B. Sumrell. "Feature Screening Using Signal-to-Noise Ratios," *submitted to Neurocomputing* (1997).
20. K.A. Greene, K.W. Bauer. "A Preliminary Investigation of Selection of EEG and Psychophysiological Features for Classifying Pilot Workload." ANNIE Conference Paper, Nov 1996.
21. K.A. Greene, K.W. Bauer, S.K. Rogers et al. *Intelligent Engineering Systems Through Artificial Neural Networks (6) ed.*. New York: ASME Press, 1996. pp. 691-697.
22. Lee, Chulhee and David Landgrebe. "Decision Boundary Feature Extraction for Non-parametric Classification," *IEEE Transactions on Systems, Man and Cybernetics*, 23(2) (March/April 1993).
23. Lippman, R.P. "An Introduction to Computing with Neural Nets," *IEEE ASSP Magazine*, 4 (1987).
24. Lyle, David L. "Proof of Ranking Equivalency." personal correspondence, Feb 1997.
25. Mayo. *Breast Cancer: New Perspectives Can Replace Unrealistic Fears*. Technical Report ISSN 0741-6254, Mayo Foundation for Medical Education and Research, October 1994.
26. McCandless, 1Lt Donald A. *Detection of Clustered Microcalcifications Using Wavelets*. MS thesis, Air Force Institute of Technology, OH, December 1995.
27. Parsons, T.W. *Voice and Speech Processing*. McGraw-Hill, 1987.
28. Polakowski, Capt. William E. *Computer Aided Diagnosis of Mammographic Masses*. MS thesis, Air Force Institute of Technology, OH, December 1995.
29. Raudys, S. and A.K. Jain. *Artificial Neural Networks and Pattern Recognition: Old and New Connections*. Elsevier, 1991.
30. Rausch, Howard. "Detecting Missiles - and Cancer Cells," *Biophotonics International* (July/August 1996).
31. Rogers, Steven K. and Dennis W. Ruck. "Feature Selection for Pattern Recognition Using Multilayer Perceptrons." *The Industrial Electronics Handbook* Innodata Publishing Services, 1996.
32. Rumelhart, D.E. and J.L. McClelland. *Parallel Distributed Processing vol 1*. Foundations MIT Press, 1986.
33. Sambur, M. "Selection of Acoustic Features for Speaker Identification," *IEEE Transactions, ASSP-23*(2) (April 1975).

34. Steppe, J. and K. Bauer. "Feature Saliency Measures." survey notes, July 1995.
35. Stewart, James A. *Non-Linear Time Series Analysis*. MS thesis, Air Force Institute of Technology, OH, March 1995.
36. Sumrell, Capt. David B. *An investigation of preliminary feature screening using signal-to-noise ratios*. MS thesis, Air Force Institute of Technology, OH, March 1996.
37. Tarr, Gregory L. *Multilayered Feedforward Neural Networks for Image Segmentation*. PhD dissertation, Air Force Institute of Technology, OH, 1991.
38. Werbos, Paul J. *Beyond Regression: New Tools for Prediction and Analysis*. PhD dissertation, Harvard University, Cambridge MA, 1974.
39. Zahirniak, Capt. Daniel R. *Characterization of Radar Signals Using Neural Networks*. MS thesis, Air Force Institute of Technology, OH, December 1990.



## *Vita*

Daniel Gregg was born in January 1968 in Pascagoula, MS. He received his Bachelor of Science in Physics from the University of South Alabama in 1990. After commissioning, he was trained for Missile Operations at Vandenberg AFB CA. He served four years as a Missile Launch Officer with the distinguished 351 Missile Wing, Whiteman AFB MO. He arrived at the Air Force Institute of Technology in August 1995.

Permanent address: 210 Autumn Dr. Saraland, AL 36571

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.				
1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE March 1997		3. REPORT TYPE AND DATES COVERED Master's Thesis
4. TITLE AND SUBTITLE Decision Boundary Analysis Feature Selection for Breast Cancer Diagnosis			5. FUNDING NUMBERS	
6. AUTHOR(S) Daniel W. Gregg, Capt. USAF				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Air Force Institute of Technology WPAFB OH 45433-6583			8. PERFORMING ORGANIZATION REPORT NUMBER AFIT/GOR/ENG/97M-04	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Wright Aeromedical Laboratories Wright-Patterson AFB, OH 45433			10. SPONSORING / MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES				
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for Public release; Distribution Unlimited			12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words)  This thesis has a three-fold contribution. First, a comparison will be made between Ruck's saliency as proposed in previous works and a new variant of Ruck's saliency to be introduced. The results of the new method will prove to be superior. Classification accuracy is improved by over 7 percentage points. Secondly, a proposal will be presented which establishes how one may use the eigenvalue/eigenvector pairs from DBFA for feature saliency. This proposal will also provide proof of the equivalence of DBFA and the Ruck variant proposed in this thesis. Because the Ruck variant is easier to calculate, it is suggested that DBFA is unnecessary. Finally, this thesis will investigate the application of classifier-free feature screening of a large feature space. A correlation-based procedure will be developed which has proven to outperform other saliency metrics such as the Fisher ratio and derivative-based techniques such as Ruck's saliency. This procedure has produced classification accuracies 10 percentage points higher than that of Fisher saliency while achieving a slightly better (2 percentage points) classification accuracy than the best derivative-based results. In addition, a combined process will be implemented which is superior to any stand alone technique. Classification results from the combined technique are 10 percentage points higher than the best results from any of the other methods. The applicability of the proposed technique is limited in this research to two-class pattern recognition problems, but may be extended to multi-class problems.				
14. SUBJECT TERMS Decision Boundary Analysis; Cascade Correlation; Feature Saliency.			15. NUMBER OF PAGES 82	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED	18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED	19. SECURITY CLASSIFICATION OF ABSTRACT UNCLASSIFIED	20. LIMITATION OF ABSTRACT UL	

## GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to *stay within the lines* to meet optical scanning requirements.

**Block 1. Agency Use Only (Leave blank).**

**Block 2. Report Date.** Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year.

**Block 3. Type of Report and Dates Covered.** State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

**Block 4. Title and Subtitle.** A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

**Block 5. Funding Numbers.** To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C - Contract	PR - Project
G - Grant	TA - Task
PE - Program Element	WU - Work Unit Accession No.

**Block 6. Author(s).** Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

**Block 7. Performing Organization Name(s) and Address(es).** Self-explanatory.

**Block 8. Performing Organization Report Number.** Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

**Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es).** Self-explanatory.

**Block 10. Sponsoring/Monitoring Agency Report Number.** (If known)

**Block 11. Supplementary Notes.** Enter information not included elsewhere such as: Prepared in cooperation with...; Trans. of...; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

**Block 12a. Distribution/Availability Statement.** Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

**DOD** - See DoDD 5230.24, "Distribution Statements on Technical Documents."

**DOE** - See authorities.

**NASA** - See Handbook NHB 2200.2.

**NTIS** - Leave blank.

**Block 12b. Distribution Code.**

**DOD** - Leave blank.

**DOE** - Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.

**NASA** - Leave blank.

**NTIS** - Leave blank.

**Block 13. Abstract.** Include a brief (*Maximum 200 words*) factual summary of the most significant information contained in the report.

**Block 14. Subject Terms.** Keywords or phrases identifying major subjects in the report.

**Block 15. Number of Pages.** Enter the total number of pages.

**Block 16. Price Code.** Enter appropriate price code (*NTIS only*).

**Blocks 17. - 19. Security Classifications.** Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

**Block 20. Limitation of Abstract.** This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.